



BIONUMERICS®

version 8 - PLUGINS



MLST online plugin

Contents

1	Starting and setting up BIONUMERICS	5
1.1	Introduction	5
1.2	Startup program	5
1.3	Creating a new database	6
2	Installing the MLST online plugin	9
2.1	MLST setup	9
2.2	The Plugin installation toolbox	10
2.3	Online MLST repositories	11
2.4	Creating a custom MLST scheme	14
2.5	MLST plugin settings	16
2.6	Updating the allele trimming patterns	17
3	Importing sequences and sequence trace files	19
3.1	Introduction	19
3.2	Import routines	19
3.3	The Assembler window	20
4	Importing allele numbers from external files	23
4.1	Introduction	23
4.2	Import routines	23
5	Alleles and profiles	25
5.1	Identifying alleles	25
5.2	Identifying allelic profiles	27
5.3	Identifying alleles and allelic profiles	28
5.4	Updating locally stored allele and profile information	28
5.5	Editing MLST information in a custom MLST database	29
5.5.1	Introduction	29
5.5.2	Editing alleles	29
5.5.3	Editing profiles	31
5.5.4	Editing clonal complexes	32
5.6	Exporting alleles and profiles	33
5.7	MLST search	33
6	MLST data analysis	35
6.1	Selections in BIONUMERICS	35
6.2	The Comparison window	35
6.3	Cluster analysis	36

NOTES

SUPPORT BY APPLIED MATHS

While the best efforts have been made in preparing this manuscript, no liability is assumed by the authors with respect to the use of the information provided.

Applied Maths will provide support to research laboratories in developing new and highly specialized applications, as well as to diagnostic laboratories where speed, efficiency and continuity are of primary importance. Our software thanks its current status for a part to the response of many customers worldwide. Please contact us if you have any problems or questions concerning the use of BIONUMERICS[®], or suggestions for improvement, refinement or extension of the software to your specific applications:

Applied Maths NV

Keistraat 120
9830 Sint-Martens-Latem
Belgium
PHONE: +32 9 2222 100
FAX: +32 9 2222 102
E-MAIL: BE-DAU-INFO@biomerieux.com
URL: <https://www.applied-maths.com>

Applied Maths, Inc.

11940 Jollyville Road, Suite 115N
Austin, Texas 78759
U.S.A.
PHONE: +1 512-482-9700
FAX: +1 512-482-9708
E-MAIL: US-DAU-INFO@biomerieux.com

LIMITATIONS ON USE

The BIONUMERICS[®] software, its plugin tools and their accompanying guides are subject to the terms and conditions outlined in the License Agreement. The support, entitlement to upgrades and the right to use the software automatically terminate if the user fails to comply with any of the statements of the License Agreement. No part of this guide may be reproduced by any means without prior written permission of the authors.

Copyright ©1998-2020, Applied Maths NV. All rights reserved.

BIONUMERICS[®] is a registered trademark of Applied Maths NV. All other product names or trademarks are the property of their respective owners.

BIONUMERICS® uses following third-party software tools and libraries:

- Python 2.7.4 release from the Python Software Foundation, <http://www.python.org/>
- Xerces library for XML input and output from the Apache Software Foundation, <https://xerces.apache.org/>
- NCBI toolkit version 2.2.28, <http://www.ncbi.nlm.nih.gov/BLAST/>
- SRA Toolkit, <https://ncbi.github.io/sra-tools/>
- Boost c++ libraries, <http://www.boost.org/>
- Samtools for interacting with SAM / BAM files, <http://www.htslib.org/download/>
- 7-Zip (7za.exe), <http://www.7-zip.org/>
- Zlib library, <https://zlib.net/>
- Pigz for parallel gzip compression, <https://zlib.net/pigz/>
- Cairo 2D graphics library version 1.12.14, <http://cairographics.org/>
- Crypto++ library version 5.5.2, <http://www.cryptopp.com/>
- OpenSSL library, <https://www.openssl.org/>
- libSVM library for Support Vector Machines, <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
- SQLite version 3.7.17, <http://www.sqlite.org/>
- pymzML Python module for high throughput bioinformatics on mass spectrometry data, <https://github.com/pymzml/pymzML>
- Numpy Python library version 1.8.1, <http://www.numpy.org/>
- BioPython Python library version 1.64, <http://www.biopython.org/>
- PIL Python library version 1.1.7, <http://www.pythonware.com/products/pil/>
- Chromium Embedded Framework, <https://bitbucket.org/chromiumembedded/cef/wiki/Home>
- SPAdes genome assembler version 3.13.1, <http://bioinf.spbau.ru/spades> *
- SKESA version 2.3.0, <https://github.com/ncbi/SKESA/releases>
- Unicycler version 0.4.8, <https://github.com/rrwick/Unicycler/releases> *
- Velvet for Windows, source code can be downloaded from <https://www.applied-maths.com/download/open-source>
- Ray for Windows, source code can be downloaded from <https://www.applied-maths.com/download/open-source>
- Bowtie2 version 2.2.5 (<http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>)*
- SNAP version 1.0.18, <http://snap.cs.berkeley.edu/>
- RAxML version 8.2.11, <https://github.com/stamatak/standard-RAxML/releases>
- FastTree version 2.1.10, <http://www.microbesonline.org/fasttree/>

- CFSAN SNP pipeline version 0.8.2, <https://github.com/CFSAN-Biostatistics/snp-pipeline> *
- Prokka version 1.12, <https://github.com/tseemann/prokka> *

*: On Calculation Engine only

Chapter 1

Starting and setting up BIONUMERICS

1.1 Introduction

This guide is designed as a tutorial for the *Multi Locus Sequence Typing plugin* of BIONUMERICS. With the *MLST online plugin* you can:

1. Download the allele numbering, sequence types and clonal complex information for a selected organism from a large number of public online repositories (e.g. PubMLST.org, MLST.net, ...) and query this information.
2. Query the databases available on the PubMLST.org (<http://pubmlst.org>) directly without using the standard web interface.
3. Store custom allele numbering, sequence type and clonal complex information in the BIONUMERICS database, and query this information.

The *MLST online plugin* is supported in the **BIONUMERICS-SEQ** and **BIONUMERICS-SUITE** configurations.

1.2 Startup program

Make sure the latest version of BIONUMERICS is installed (<https://www.applied-maths.com/download/software>). The installation manual can be downloaded from <https://www.applied-maths.com/download/manuals>.

When BIONUMERICS is launched from the Windows start panel or when the BIONUMERICS shortcut () on your computer's desktop is double-clicked, the **Startup program** is run. This program shows the *BIONUMERICS Startup* window (see Figure 1.1).

A new BIONUMERICS database is created from the Startup program by pressing the  button.

An existing database is opened in BIONUMERICS with  or by simply double-clicking on a database name in the list.

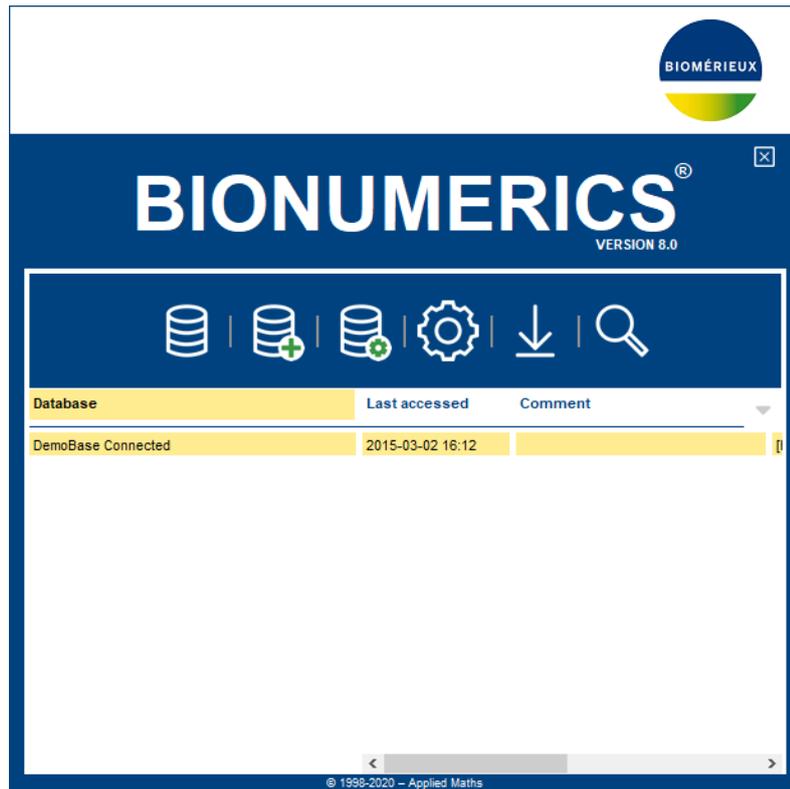


Figure 1.1: The *BIONUMERICS* Startup window.

1.3 Creating a new database

3.1 Press the  button in the *BIONUMERICS* *BIONUMERICS* Startup window to enter the *New database* wizard.

3.2 Enter a name for the database, and press **<Next>**.

A new dialog box pops up, prompting for the type of database (see Figure 1.2).

3.3 Leave the default option selected and press **<Next>**.

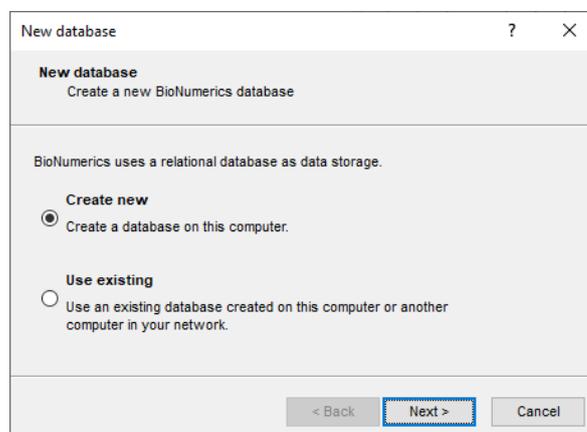


Figure 1.2: The *New database* wizard page.

A new dialog box pops up, prompting for the database engine (see Figure 1.3).

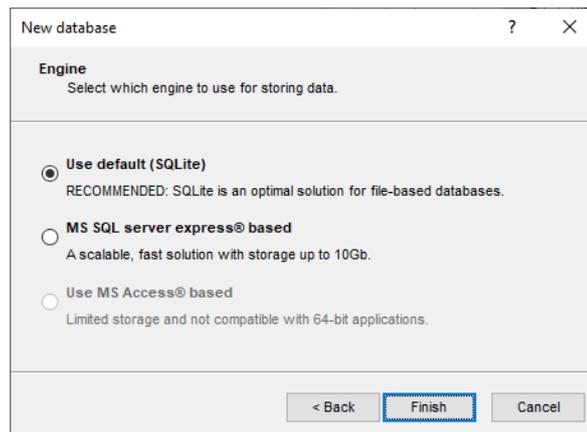


Figure 1.3: The *Database engine* wizard page.

3.4 Leave the default option selected and press <**Finish**> to complete the setup of the new database.

Chapter 2

Installing the MLST online plugin

2.1 MLST setup

Before an organism can be typed with MLST, an *MLST scheme* needs to be defined for that organism. An MLST scheme consists of a list of loci (typically seven), which are sequenced and assigned an allele number. Furthermore, in order to achieve an unambiguous nomenclature, a list of *allele variants* per locus needs to be provided and a list of *allelic profiles* corresponding to *sequence types*.

The *MLST online plugin* can be set up in different ways, depending on the available information for the organism of interest. An MLST scheme is defined by its **profile and allele files**: a tab-delimited text file with the sequence types and their corresponding allelic profiles and, for each of the loci, a FASTA-formatted text file with the allele variant definitions. These files can be made available on the local computer, the network or on the internet. A path or URL should be specified for each of the files. Even if no MLST scheme is publicly available, the plugin can be used to type organisms using MLST.

When a MLST data repository is available online for the organism of your interest, the MLST setup comes down to simply picking the organism from the provided list. The profile and allele variant definitions will be downloaded and stored locally in the BIONUMERICS database. This information can be automatically updated from the online definition files.

It can occur that your organism is not listed (meaning that it is not available from a known online data repository), but nevertheless an MLST scheme has been published for this organism and profile and allele files are available. In this case, the organism name, locus names and the locations of the profile and allele files (a path on the local computer, network drive or URL) should be entered manually. The profile and allele definitions will then be downloaded and can be automatically updated from the definition files.

In case no MLST scheme has been published yet for the organism of your interest, the only option is to store all data locally and to update the profile and allele definitions manually.

Once the *MLST online plugin* is installed, it will not be possible anymore to modify the MLST scheme (organism name, locus names) for that database. It is possible, however, to change the way the MLST data is updated. For example, the following scenario can be accommodated for:

1. Startup of a new MLST scheme: all data is stored locally; manual access only
2. Share the MLST scheme between different databases via local files (tools are available to export the profile and allele definitions to text files)
3. Make the MLST scheme public by putting the definition files on the web

The following paragraphs will guide you through the *MLST online plugin* installation process.

2.2 The Plugin installation toolbox

The *Plugins* dialog box can be called from the *Main* window by selecting **File > Install / remove plugins...** (see Figure 2.1).

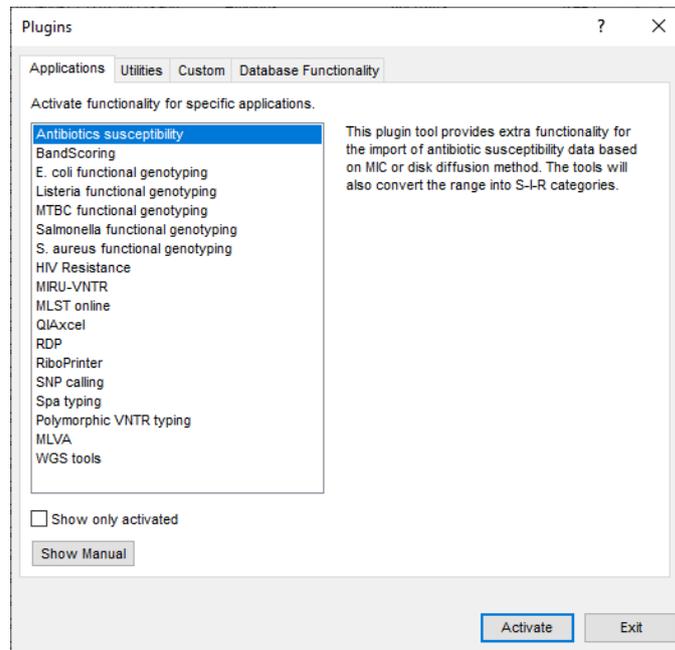


Figure 2.1: The *Plugins* dialog box.

When a particular plugin is selected from the list of plugins, a short description appears in the right panel.

A selected plugin can be installed with the **<Activate>** button. The software will ask for confirmation before installation. Some plugins are only supported in specific BIONUMERICS configurations. If the plugin is not supported by your BIONUMERICS configuration, it cannot be installed and an error message will be generated.

Once a plugin is installed, it is marked with a green V-sign. It can be removed again with the **<Deactivate>** button.

If the selected plugin is documented, pressing **<Show Manual>** will open its manual in the *Help* window.

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

BIONUMERICS asks the user to confirm the installation of the *MLST online plugin* (see Figure 2.2). Installation of the plugin requires administrator privileges on the relational database.

- 2.2 Press **<Yes>** to start with the installation.

In the first step of the wizard, the user is asked to select the organism source (see Figure 2.3).

- **Select organism from on-line list:** Allele numbering, sequence type and clonal complex information from online repositories is used for typing. All available organisms are listed in the next step of the wizard (see 2.3).

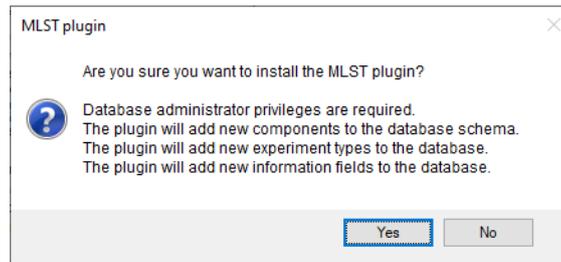


Figure 2.2: Confirmation message that appears when installing the *MLST online plugin*.

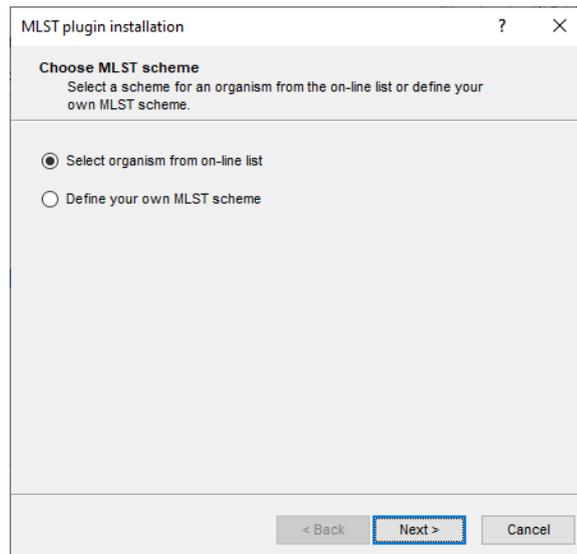


Figure 2.3: The *Choose MLST scheme* dialog box.

- **Define your own MLST scheme:** Custom information is stored in the BIONUMERICUS database (see 2.4).

2.3 Online MLST repositories

In the *Choose MLST scheme* dialog box choose the option **Select organism from on-line list** and press **<Next>** to call the *Select an organism* dialog box (see Figure 2.4).

Any organism for which an MLST repository is available online, is listed in the *Select an organism* dialog box. For the selected organism, abbreviations of the housekeeping genes are listed in the right panel, together with the URL where the database is located and the number of available profiles.

Pressing **<Next>** calls the *Define profile and allele files* dialog box (see Figure 2.5).

In the *Define profile and allele files* dialog box, the location of the profile and allele definition files are shown. For public MLST schemes, the default locations point to the correct file URLs.

It is possible to specify a different location for the profile and allele definition files, which can be located on your own computer, local area network or on the internet. This is however unnecessary in a standard setup.

Checking the option **Update profiles and alleles at database startup** ensures that the profile and

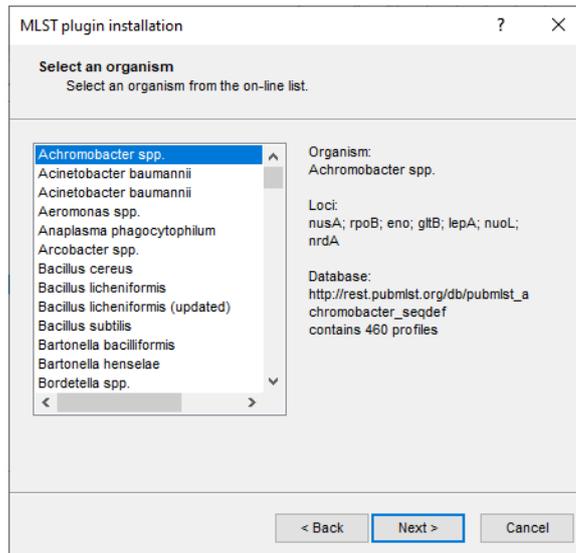


Figure 2.4: The *Select an organism* dialog box.

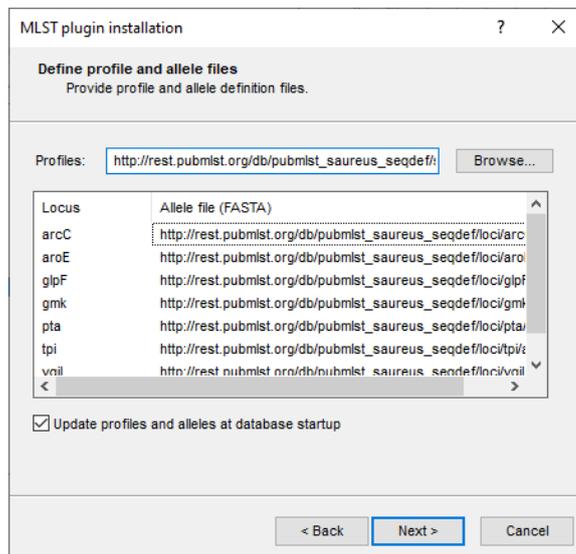


Figure 2.5: The *Define profile and allele files* dialog box.

allele definitions are always up-to-date, by updating the latter each time when the BIONUMERICs database is opened.

All remotely stored MLST information (allele numbering, sequence types and clonal complexes) for the selected organism will be downloaded and stored in the BIONUMERICs database during plugin installation.

Pressing **<Next>** calls the *Allele trimming patterns* dialog box (see Figure 2.6).

In the *Allele trimming patterns* dialog box the start and stop trimming patterns can be entered manually for each of the alleles, or one can specify to **Calculate trimming patterns automatically** (recommended). If the trimming patterns are automatically calculated, a **Trim pattern length** (default 20 bases) can be set. The trim patterns can be recalculated automatically each time the database is loaded by checking **Update trimming patterns at database startup**.

Pressing **<Next>** calls the *Database info fields* dialog box (see Figure 2.7).

Figure 2.6: The *Allele trimming patterns* dialog box.

Figure 2.7: The *Database info fields* dialog box.

In the *Database info fields* dialog box, the program prompts for database information fields to store the **Sequence types** and **Clonal complexes** information. One can either choose the default fields, select existing fields or enter new field names. When manually entering new field names, please note that information field names cannot start with a space.

Pressing **<Next>** calls the final dialog in the wizard (see Figure 2.8).

Pressing **<Finish>** in the *Finish installation* dialog box starts with the installation of the *MLST online plugin*.



In case of large MLST databases available online, the installation may take several minutes, depending on the speed of your internet connection.

When the *MLST online plugin* is successfully installed, a confirmation message is displayed. Close and reopen the database to activate the features of the *MLST online plugin*.

The *MLST online plugin* installs menu items in the main menu of the software under **MLST** (see

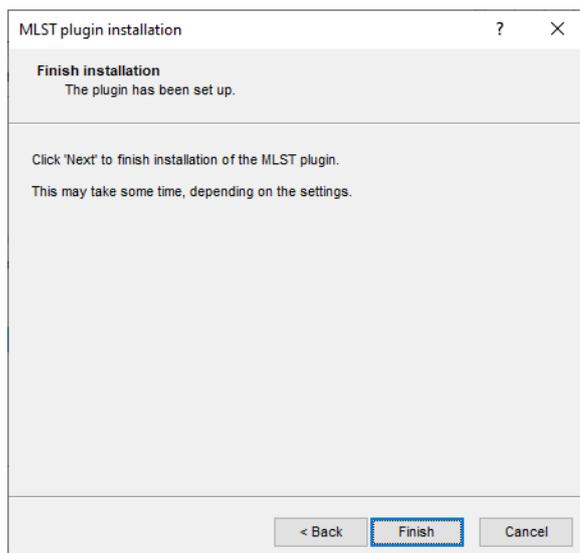


Figure 2.8: The *Finish installation* dialog box.

Figure 2.9) and in the *Contig assembly* window.

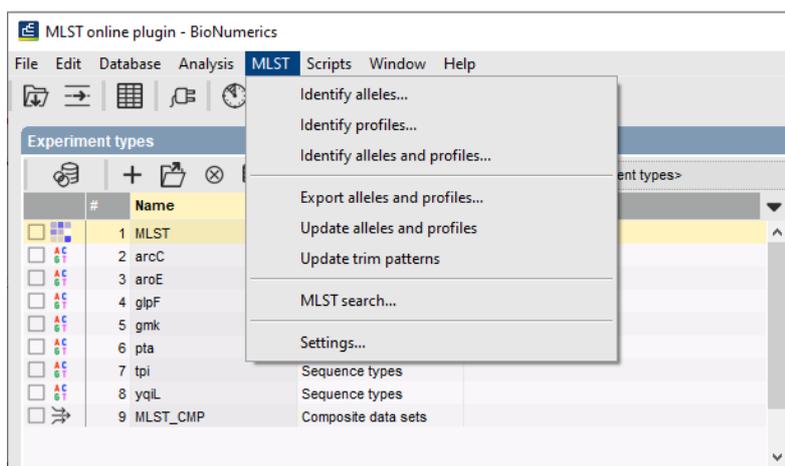


Figure 2.9: MLST menu items in the *Main* window.

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 typically seven sequence types, each named after a housekeeping gene.

2.4 Creating a custom MLST scheme

In case no MLST scheme is available yet for the organism of your interest, the *MLST online plugin* allows you to create a custom MLST scheme.

In the *Choose MLST scheme* dialog box (see Figure 2.3), select **Define your own MLST scheme** and press <**Next**> to call the *Define your own MLST scheme* dialog box (see Figure 2.11).

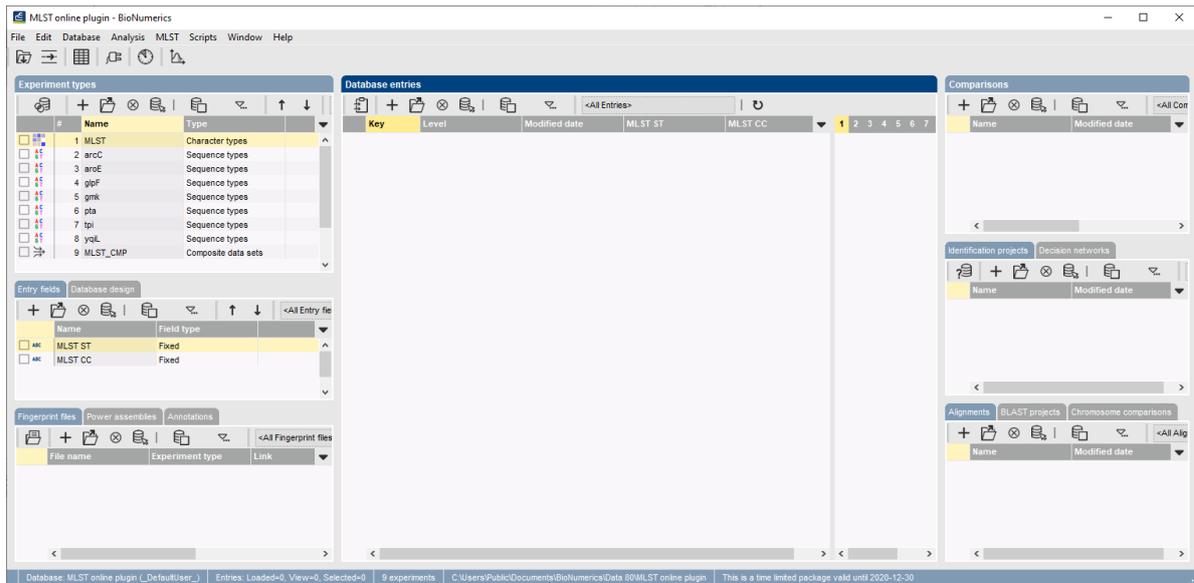


Figure 2.10: The Main window after installation of the MLST online plugin.

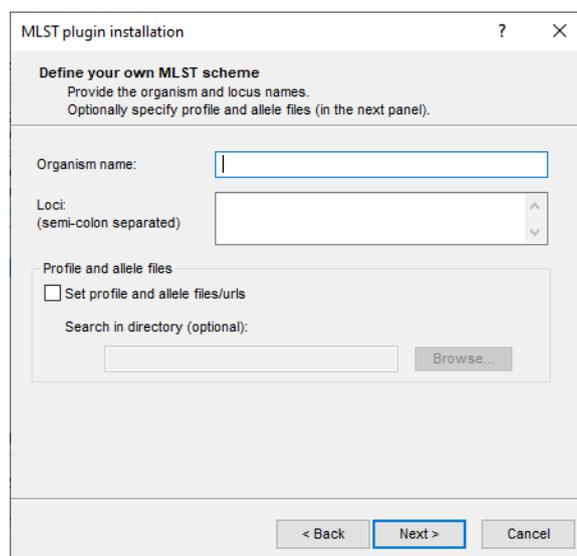


Figure 2.11: The Define your own MLST scheme dialog box.

In the *Define your own MLST scheme* dialog box, you are asked to provide the **Organism name** and the **Loci** that were sequenced. The loci should be entered, separated by semi-colons.

In case text files containing profile and allele information are already available (e.g. if they were sent to you by a colleague who works on the same organism), you should check **Set profile and allele files/URLs**. With this option checked, pressing **<Next>** will take you to the *Select an organism* dialog box where profile and allele file locations can be specified (see Figure 2.5), which will initially be empty.

When **Set profile and allele files/URLs** is unchecked, the next step will be dealing with the trimming positions.

When FASTA text files for allele definitions are used, they should be formatted as follows:

```
> [ID]
```

ACTG...

With [ID] an integer allele identifier (without brackets). Alternatively, the identifier line could also be: ">_[ID] ", ">-[ID] ", ">[LOCUS] [ID] ", ">[LOCUS] _[ID] " or ">[LOCUS]-[ID]".

When the option **Set profile and allele files/URLs** was checked, specify the profile and allele file locations in the *Select an organism* dialog box.

When the option **Set profile and allele files/URLs** was unchecked, enter the **Start pattern** and **Stop pattern** for all loci or check **Calculate trimming patterns automatically** in the *Allele trimming patterns* dialog box.

Proceed with the steps explained in 2.3 to complete the installation of the *MLST online plugin*.

2.5 MLST plugin settings

All settings that were entered during installation of the MLST online plugin, except organism name and loci, can be changed later on. In the *Main* window, select **MLST > Settings** to call the *MLST plugin settings* dialog box (see Figure 2.12).

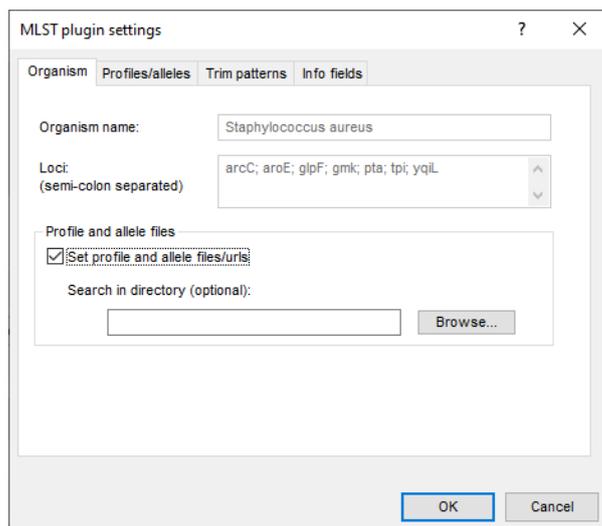


Figure 2.12: The *MLST plugin settings* dialog box, *Organism* tab.

The *MLST plugin settings* dialog box consists of four different tabs: the *Organism*, *Profiles/alleles*, *Trim patterns*, and *Info fields* tab. To access the relevant settings, simply click on the corresponding tab in the top part of the dialog box.

In the *Organism* tab, which is displayed initially, the **Organism name** and **Loci** are read-only, since the MLST scheme cannot be changed after installation. Checking **Set profile and allele files/URLs** will update profile and allele definitions from external text files, which can be located on your computer, on a network computer or on the internet. Optionally, a directory can be entered or browsed for.

The *Profiles/alleles* tab basically contains the same settings as the *Define profile and allele files* dialog box (see Figure 2.5). The location where the program looks for the profile and/or allele files can be edited. This action is of course only necessary when the physical location of these files on your computer or on the (web) server has changed.

In the *Trim patterns* tab, the **Start pattern** and **Stop pattern** for each of the loci can be edited (only if **Calculate trimming patterns automatically** is unchecked), similar as in the *Allele trimming*

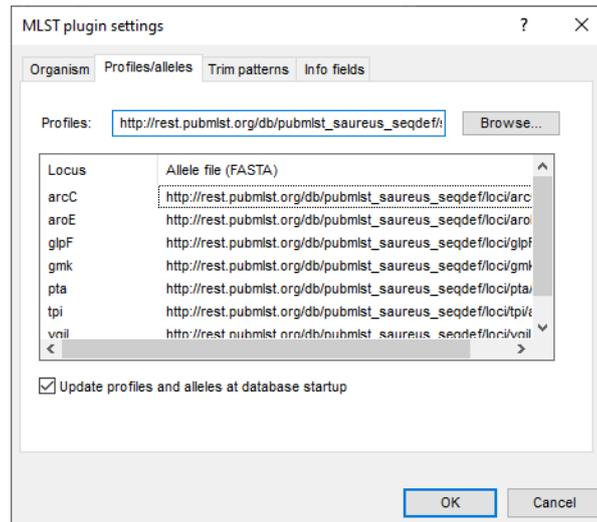


Figure 2.13: The *MLST plugin settings* dialog box, *Profiles/alleles* tab.

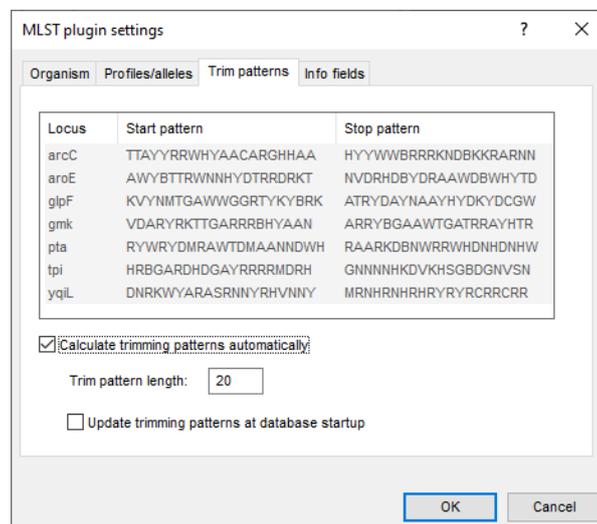


Figure 2.14: The *MLST plugin settings* dialog box, *Trim patterns* tab.

patterns dialog box (see Figure 2.6). Also, it can be specified to update the trimming patterns each time that the database is loaded. A **Trim pattern length** can be entered if the trimming patterns are calculated automatically.

From the *Info fields* tab, different information can be selected for storage of **Sequence types** and **Clonal complexes** information. The functionality of this tab corresponds to the *Database info fields* dialog box (see Figure 2.7).

2.6 Updating the allele trimming patterns

All trimming patterns are displayed in the *Trim patterns* tab of the *MLST plugin settings* dialog box (see 2.5). In the *Trim patterns* tab, the **Start** and **Stop patterns** for each of the loci are displayed. These patterns will be automatically displayed in the *Assembly trimming settings* dialog box when importing sequences into the database with the batch import routine.

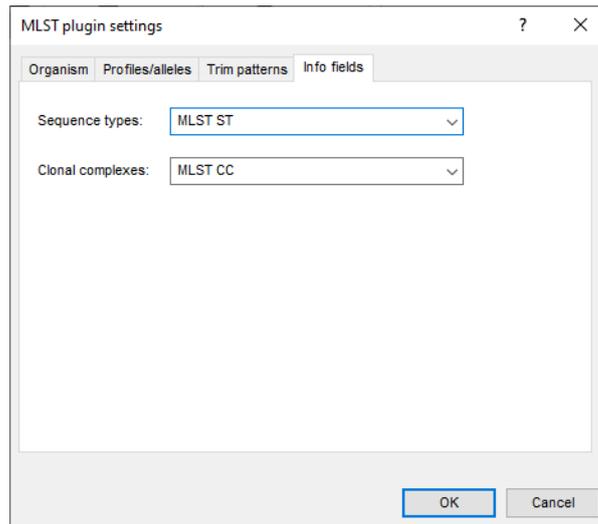


Figure 2.15: The The *MLST plugin settings* dialog box, *Info fields* tab.

When the option ***Update trimming patterns at database startup*** is checked in the *Trim patterns* tab, the trim patterns are automatically updated each time the database is loaded.

To update the trim patterns of the selected organism manually, use the ***MLST > Update trim patterns*** command.

Chapter 3

Importing sequences and sequence trace files

3.1 Introduction

Sequences and sequence trace files can be imported (and assembled) in BIONUMERICS and stored in the corresponding housekeeping gene sequence type experiments using the import routines available in the *Import* dialog box. After import, the (consensus) sequences can be screened against the downloaded MLST sequences (see 5.1 for more information).

3.2 Import routines

The *Import* dialog box is called with the command **File** > **Import...** (📁, **Ctrl+I**). The import tree options are organized in groups based upon the type of data. Sequence data can be imported in BIONUMERICS in several ways (see Figure 3.1).

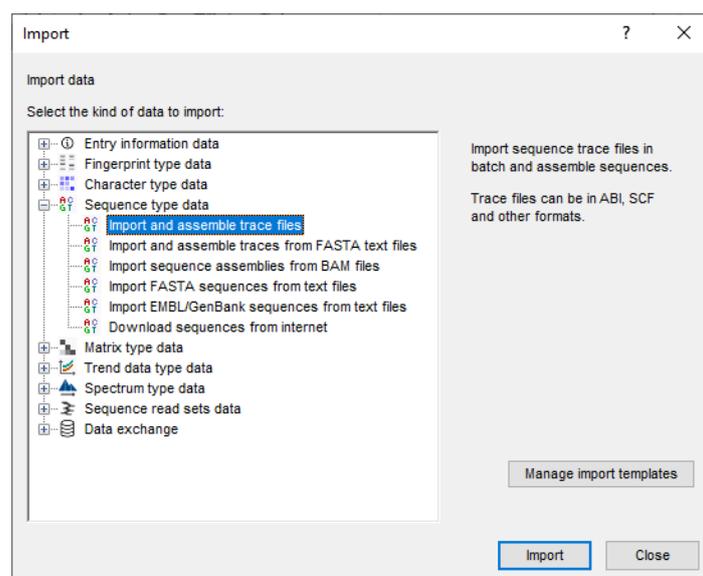


Figure 3.1: The sequence import routines in the Import tree.

The import routines will not be covered in detail in this manual. More detailed information can be

found in the BIONUMERICS manual.

3.3 The Assembler window

When importing and assembling sequence trace files in BIONUMERICS, the consensus sequence can be screened for (nearest) allele matches with **MLST > Identify allele** in the *Contig assembly* window.

The *MLST identification plot window* will pop up. The sequence type of the sequence that is shown in Assembler is displayed in white. The other sequence types are shown in gray (see Figure 3.2). The identification result is shown below the sequence type boxes.

In case the trimmed consensus sequence could not be matched to one of the alleles (e.g. due to the presence of unresolved bases), a table is shown in the lower left part of the *MLST identification plot window* displaying the best matched alleles in the **Allele ID** column and number of mismatches with the consensus sequence in the **Mismatches** column (see Figure 3.2 a). Matched allele IDs with a maximum of 5 mismatches are reported. All allele IDs are clickable. A second table is displayed in the right part of the window listing the sequence differences between the consensus sequence and the selected allele ID in the left table. The base and base position on the consensus sequence are displayed in the **Contig** and **Assembly** columns, the corresponding bases on the matched allele are displayed in the **Allele** column. Clicking on one of the editing suggestions in the right table automatically updates the focus in the *Contig assembly* window. In case no match could be found after introduction of 5 or less mismatches, the text "No matching alleles found" is displayed (see Figure 3.2 b).

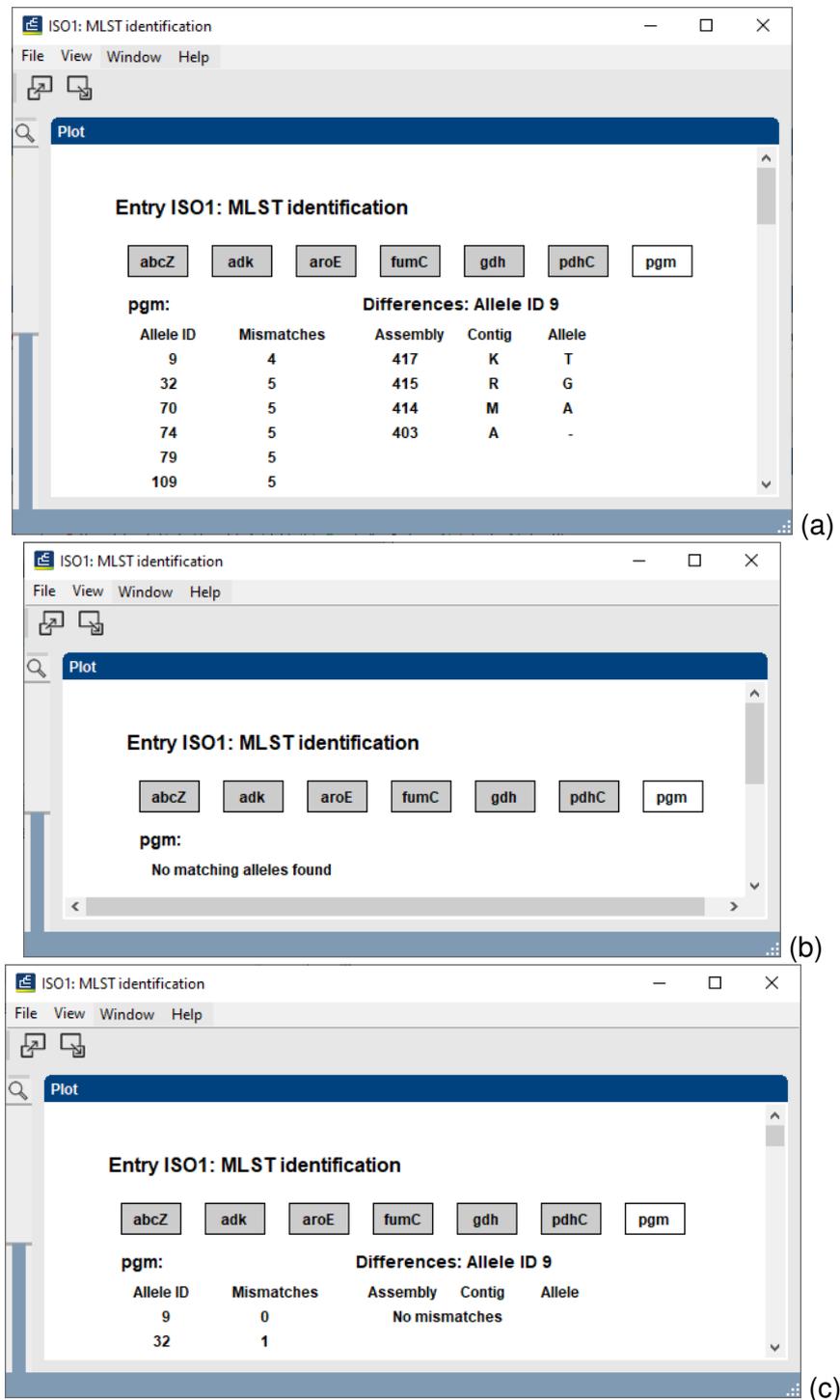


Figure 3.2: The *MLST identification plot window*: (a) Best match with allele ID 9; (b) No matching alleles; (c) Perfect match with allele ID 9.

Chapter 4

Importing allele numbers from external files

4.1 Introduction

When the MLST allele numbers are already stored in external text, Excel, or other ODBC-compatible file, these allele numbers can be imported in BIONUMERICS and stored in the character type experiment **MLST** using import routines available in the *Import* dialog box. After import, the allelic profiles can be screened against the sequence type and clonal complex information (see 5.2 for more information).

4.2 Import routines

The *Import* dialog box is called with the command **File > Import...** (📄, **Ctrl+I**). The import tree options are organized in groups based upon the type of data (see Figure 4.1).

- With the ***Import fields and characters (text file)*** option, listed under the topic ***Character type data*** in the Import tree, the allele numbers can be imported from text files in the database and linked to new or existing database entries.
- With the ***Import fields and characters (Excel file)*** option, listed under the topic ***Character type data*** in the Import tree, the allele numbers can be imported from an Excel file in the database and linked to new or existing database entries.
- With the ***Import fields and characters (ODBC)*** option, listed under the topic ***Character type data*** in the Import tree, the allele numbers can be imported from ODBC-compatible files in the database and linked to new or existing database entries.

The import routines will not be covered in detail in this manual. More detailed information can be found in the BIONUMERICS manual.

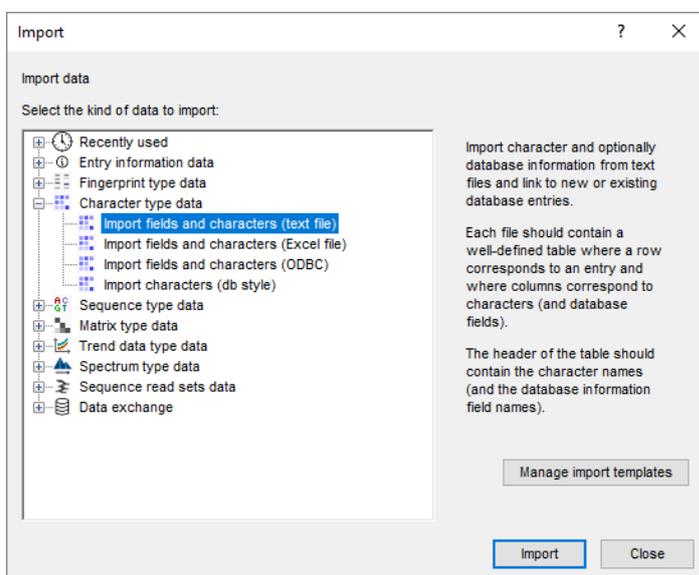


Figure 4.1: The character import routines in the Import tree.

Chapter 5

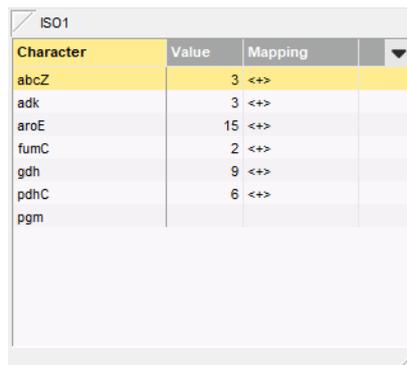
Alleles and profiles

5.1 Identifying alleles

After clean-up of the assemblies, the consensus sequences can be screened against the allele information with **MLST > Identify alleles**.

Screening can be done for all entries present in the database, or for any selection of entries in database. A selection can be made in the database using the **Ctrl** and **Shift** keys. To select all entries in the database at once, use the shortcut **Ctrl+A**. Check boxes for selected entries are indicated as . If no selection is present in the database, a message pops up prompting to confirm to run the tool on all entries in the database. Pressing **<OK>** starts the allele identification for all entries that are present in the database.

The matched allele IDs are stored in the corresponding character fields of the **MLST** character type. When a consensus sequence does not match one of the alleles in the database, the character value is left empty. Clicking on the colored dot in the **MLST** column of the *Experiment presence* panel opens the character *Experiment card* window for an entry (see Figure 5.1). The experiment card can be closed again by clicking in the small triangle-shaped button in the left upper corner.



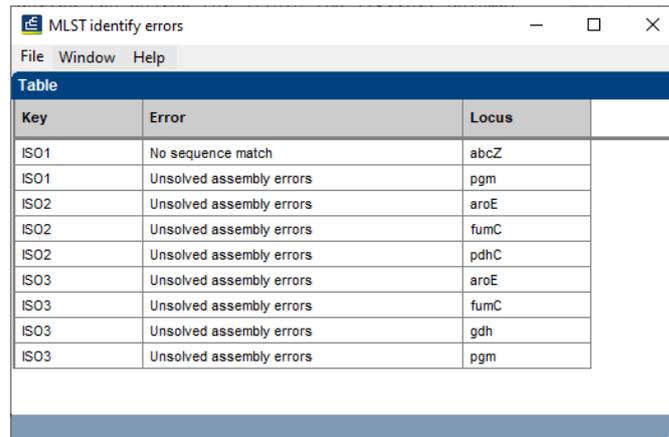
Character	Value	Mapping
abcZ	3	<+>
adk	3	<+>
aroE	15	<+>
fumC	2	<+>
gdh	9	<+>
pdhC	6	<+>
pgm		

Figure 5.1: An MLST character card, displaying allele IDs.

If one or more sequences could not be assigned to an allele ID, an *MLST error report window* pops up (see Figure 5.2 for an example). The report window displays all errors encountered during the search for allele IDs ('Error' column), the linked sequence type ('Locus' column) and the entry key ('Key' column).

The error messages can be:

- **Unresolved assembly errors:** When the status of the assembly is either "error" or "read",



Key	Error	Locus
ISO1	No sequence match	abcZ
ISO1	Unsolved assembly errors	pgm
ISO2	Unsolved assembly errors	aroE
ISO2	Unsolved assembly errors	fumC
ISO2	Unsolved assembly errors	pdhC
ISO3	Unsolved assembly errors	aroE
ISO3	Unsolved assembly errors	fumC
ISO3	Unsolved assembly errors	gdh
ISO3	Unsolved assembly errors	pgm

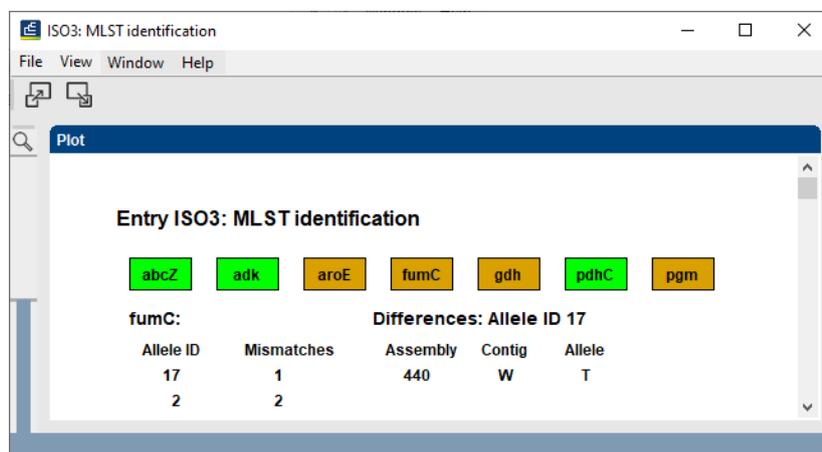
Figure 5.2: The *MLST error report window*, after identifying alleles.

i.e. when the assembly errors were not yet manually corrected or when the errors have been corrected but the status has not been changed to "solved" afterwards.

- **No sequence match:** The consensus sequence could not be matched to an existing allele ID.

Double-clicking on one of the error messages in the *MLST error report window* opens the *MLST identification plot window* of the linked entry (see Figure 5.3).

The sequence types that are assigned to an allele ID are displayed in green. The sequence types of the sequences that could not be linked to an allele ID are displayed in orange (see Figure 5.3).



Entry ISO3: MLST identification

abcZ adk aroE fumC gdh pdhC pgm

fumC: Differences: Allele ID 17

Allele ID	Mismatches	Assembly	Contig	Allele
17	1	440	W	T
2	2			

Figure 5.3: The *MLST identification plot window*: No perfect match with an allele, best match with allele ID 17.

The identification result of the selected sequence type is shown below the sequence plot:

In case the trimmed consensus sequence could not be matched to one of the alleles (e.g. due to the presence of unresolved bases) a table is shown in the lower left part of the *MLST identification plot window* displaying the best matched allele IDs in the "Allele ID" column, and number of mismatches with the consensus sequence in the "Mismatches" column (see Figure 5.3). Matched allele IDs with a maximum of 5 mismatches are reported. All allele IDs can be clicked on, in which case a second table is displayed in the right part of the window, displaying the sequence differences between the consensus sequence and the selected allele ID in the left table. The base and

base position on the consensus sequence are displayed in the "Contig" and "Assembly" columns respectively. The corresponding bases on the matched allele are displayed in the "Allele" column. Clicking on one of the sequence differences in the right table, automatically updates the focus in the *Contig assembly* window. In case no match could be found after introduction of 5 or less mismatches, the text "No matching alleles found" is displayed.

Clicking on another sequence type field automatically updates the identification results in the lower part of the *MLST identification plot window*.

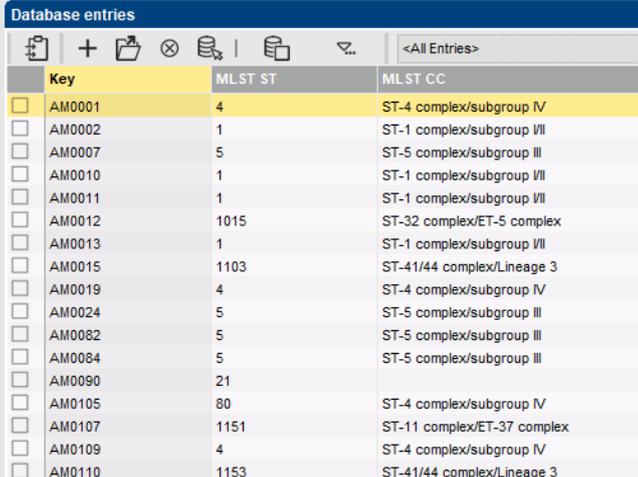
Clean up all assembly errors in the *MLST error report window* and select **MLST > Identify alleles** to update the allelic profiles for the selected entries in the database.

5.2 Identifying allelic profiles

The allelic profiles of the entries in the database are screened against the downloaded sequence type and clonal complex information with the command **MLST > Identify profiles**.

Screening can be done for all entries present in the database, or for any selection of entries in database. A selection can be made in the database using the **Ctrl-** and **Shift-**keys. To select all entries in the database at once, use the shortcut **Ctrl+A**. Check boxes for selected entries are indicated as . If no selection is present in the database, a message pops up prompting to confirm to run the tool on all entries in the database. Pressing **<OK>** starts the profile screening for all entries in the database.

The matched sequences types and clonal complexes are displayed in the MLST information fields (default **MLST ST** and **MLST CC** respectively, see Figure 5.4).



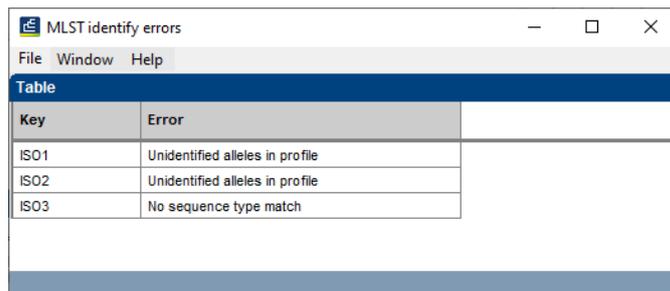
Key	MLST ST	MLST CC
<input checked="" type="checkbox"/> AM0001	4	ST-4 complex/subgroup IV
<input type="checkbox"/> AM0002	1	ST-1 complex/subgroup VII
<input type="checkbox"/> AM0007	5	ST-5 complex/subgroup III
<input type="checkbox"/> AM0010	1	ST-1 complex/subgroup VII
<input type="checkbox"/> AM0011	1	ST-1 complex/subgroup VII
<input type="checkbox"/> AM0012	1015	ST-32 complex/ET-5 complex
<input type="checkbox"/> AM0013	1	ST-1 complex/subgroup VII
<input type="checkbox"/> AM0015	1103	ST-41/44 complex/Lineage 3
<input type="checkbox"/> AM0019	4	ST-4 complex/subgroup IV
<input type="checkbox"/> AM0024	5	ST-5 complex/subgroup III
<input type="checkbox"/> AM0082	5	ST-5 complex/subgroup III
<input type="checkbox"/> AM0084	5	ST-5 complex/subgroup III
<input type="checkbox"/> AM0090	21	
<input type="checkbox"/> AM0105	80	ST-4 complex/subgroup IV
<input type="checkbox"/> AM0107	1151	ST-11 complex/ET-37 complex
<input type="checkbox"/> AM0109	4	ST-4 complex/subgroup IV
<input type="checkbox"/> AM0110	1153	ST-41/44 complex/Lineage 3

Figure 5.4: The *Main* window after screening for sequence type and clonal complex information.

If one or more allelic profiles could not be assigned to a sequence type (or clonal complex), an *MLST error report window* pops up (see Figure 5.5 for an example). The report window displays the errors encountered during the search for sequence types and clonal complexes. The errors are shown in the 'Error' column, and the linked entry key is displayed in the 'Key' column.

The error messages can be:

- **Unidentified alleles in profile:** One or more alleles are not assigned to an allele ID.



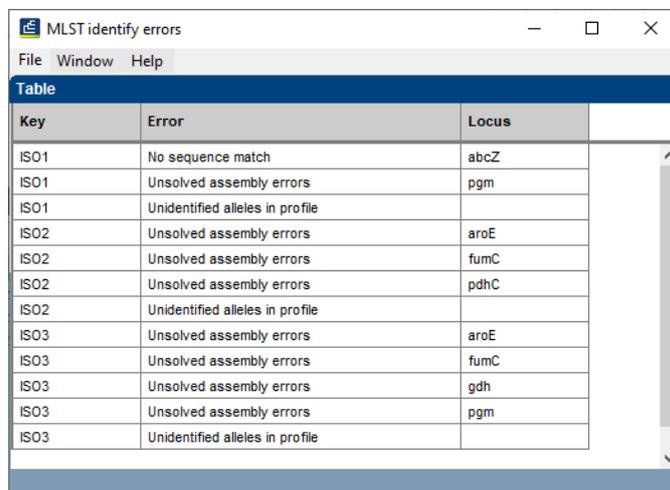
Key	Error
ISO1	Unidentified alleles in profile
ISO2	Unidentified alleles in profile
ISO3	No sequence type match

Figure 5.5: The *MLST error report window*, after identifying profiles.

- **No sequence type match:** The allelic profile could not be matched to an existing allelic profile in the locally stored database.

5.3 Identifying alleles and allelic profiles

A combined screening for alleles and profiles can be launched with the command ***MLST > Identify alleles and profiles***. This command combines the actions ***MLST > Identify allele*** and ***MLST > Identify profiles*** as described in 5.1 and 5.2, respectively. If one or more sequences could not be assigned to an allele ID and/or if one or more allelic profiles could not be assigned to a sequence type (or clonal complex), an *MLST error report window* pops up, which displays all errors encountered during the search for allele IDs, sequence types, and clonal complexes (see Figure 5.6 for an example).



Key	Error	Locus
ISO1	No sequence match	abcZ
ISO1	Unsolved assembly errors	pgm
ISO1	Unidentified alleles in profile	
ISO2	Unsolved assembly errors	aroE
ISO2	Unsolved assembly errors	fumC
ISO2	Unsolved assembly errors	pdhC
ISO2	Unidentified alleles in profile	
ISO3	Unsolved assembly errors	aroE
ISO3	Unsolved assembly errors	fumC
ISO3	Unsolved assembly errors	gdh
ISO3	Unsolved assembly errors	pgm
ISO3	Unidentified alleles in profile	

Figure 5.6: The *MLST error report window*, after identifying alleles and profiles.

5.4 Updating locally stored allele and profile information

Locally stored information is used when querying for allele, sequence type and clonal complex information in BIONUMERICs. This information can be obtained from profile and allele definition files, located on your computer, local area network or on the internet (see Figure 2.5). With the command ***MLST > Update alleles and profiles***, the locally stored MLST information in the BIONUMERICs database is then compared with the information in the allele and profile files and

updated when new information is found. A warning is generated if existing alleles or profiles are changed or deleted. When the option **Update profiles and alleles at database startup** is checked, the MLST information is automatically updated each time the database is loaded. In case profile and allele files are not available, the locally stored MLST information should be manually updated (see 5.5).

5.5 Editing MLST information in a custom MLST database

5.5.1 Introduction

The *locally* stored MLST information of a *custom* MLST database can be updated from profile and allele definition files specified in the *Profiles/alleles* tab (see 5.4).

If no allele and profile definition files are specified, the MLST information can be updated manually using the *Update MLST alleles/profiles* dialog box. Selecting **MLST > Update alleles and profiles** calls this dialog (see Figure 5.7). The *Update MLST alleles/profiles* dialog box consists of three different tabs: the *Alleles*, *Profiles* and *Clonal Complexes* tab. To access the relevant settings, simply click on the corresponding tab in the top part of the dialog box.

5.5.2 Editing alleles

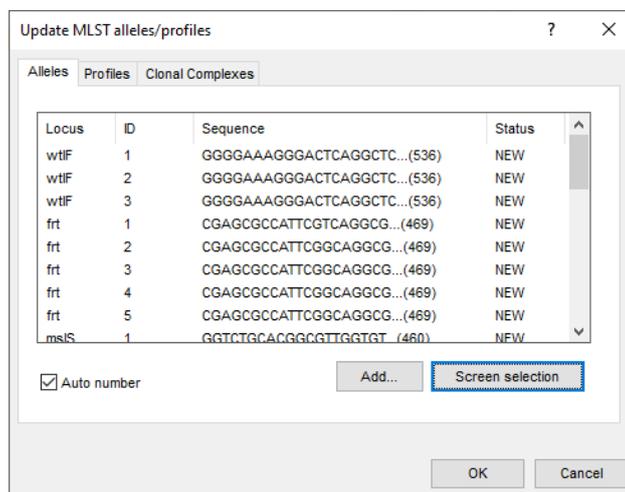


Figure 5.7: Allele information in the custom MLST database.

In the *Alleles* tab, sequences can be added to the database.

When pressing the **<Screen selection>** button, the software compares all sequences of the current entry selection with the sequences that are already present in the custom database.

All newly detected sequences are added to the allele list in the *Alleles* tab. The **Locus** column displays the name of the sequence type, the 'Sequence' field displays the first nucleotides of the allele sequence, and the last column displays the **Status** information. If the option **Auto number** was unchecked, no IDs are assigned to the newly added sequences in the **ID** column. An ID can be added to a sequence by double-clicking on a sequence in the list. If the **Auto number** option was checked, the software automatically assigns unique IDs to the newly added sequences (recommended).

Double-click on an allele in the list if you want to edit the allele information. This calls the *Edit allele* dialog box.

Instead of letting the program screen a selection of entries for alleles that are not yet present in the database, a new allele can also be added manually by pressing the <**Add**> button. This calls the *Edit allele* dialog box.

Figure 5.8: Adding an allele to the custom MLST database.

Select the sequence type from the **Locus** drop-down menu and enter the sequence in the **Allele Sequence** text box. If the option **Auto number** is unchecked in the *Alleles tab*, specify a unique allele number in the **Allele ID** text box. If the **Auto number** option is checked, the **Allele ID** text box is not editable and the software will automatically assign a new ID to the new allele when the allele is added to the database.

Pressing <**OK**> in the *Update MLST alleles/profiles* dialog box adds all new alleles in the list to the database.



It is possible to view the alleles or allelic profiles stored in the database with an *object query*. In the *Main* window, choose **Database > Object queries...** (☰) and select "<Create new>" from the drop-down menu that appears. As **Object to report**, select "MLST - Alleles" or "MLST - Sequence types" and press <**OK**>.

The update alleles and profiles function furthermore allows you to edit and delete alleles that are already stored in the database. An object query (see above) can be used to retrieve the allele ID of the allele to be edited.

In the *Alleles tab*, uncheck **Auto number** and press the <**Add**> button to call the *Edit allele* dialog box (see Figure 5.8). Select the sequence type from the **Locus** drop-down menu and enter the **Allele ID** of the allele to be edited.

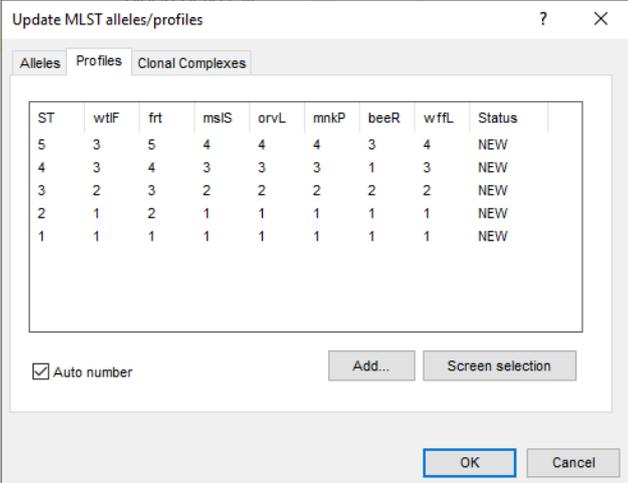
To *edit* the allele, paste a corresponding sequence in the **Allele Sequence** text box and press <**OK**>. A confirmation message will appear, confirming the change. If you press <**Yes**>, the edited allele will be listed as "UPDATE" in the allele list. The allele will effectively be edited in the database after pressing the <**OK**> button in the *Update MLST alleles/profiles* dialog box.

To *delete* the allele from the database, leave the **Allele Sequence** text box empty and press <**OK**>. A confirmation message will appear, confirming the removal of the allele. If you press <**Yes**>, the allele will be listed as "DELETE" in the allele list. The allele will effectively be deleted from the database after pressing the <**OK**> button in the *Update MLST alleles/profiles* dialog box.

Please note that adding new alleles to the database is also possible from within Assembler. When identifying an allele in the *Contig assembly* window, new alleles might be found, i.e. no matching

allele is detected in the database, while no errors or inconsistencies are present in the alignment. In that case, you can select **MLST > Add to alleles** in the *Contig assembly* window to add the allele to the database with an automatically assigned number. When the allele indeed does not yet exist in the database, a confirmation message will appear. An error will be generated in case the allele was already present in the database.

5.5.3 Editing profiles



ST	wtIF	frt	msIS	orvL	mnkP	beeR	wffL	Status
5	3	5	4	4	4	3	4	NEW
4	3	4	3	3	3	1	3	NEW
3	2	3	2	2	2	2	2	NEW
2	1	2	1	1	1	1	1	NEW
1	1	1	1	1	1	1	1	NEW

Figure 5.9: Profile information in the custom MLST database.

In the *Profiles tab*, profile information can be added to the database.

When pressing the **<Screen selection>** button, the software compares the allelic profiles of the current entry selection to the profiles that are already present in the custom database. All newly detected complete profiles are added to the profiles list.

The **ST** column displays the name of the sequence type; the last column displays the **Status** information. If the option **Auto number** was unchecked in the *Profiles tab*, no sequence types are assigned to the newly added profiles in the list. A ST can be added by double-clicking on a profile in the list. If the **Auto number** option was checked, the software automatically assigns unique ST IDs to the newly added profiles.

Double-click on a profile in the list if you want to edit the profile information. This pops up the *Edit profile* dialog box.

A new profile can also be added manually by pressing the **<Add>** button. This calls the *Edit profile* dialog box.

Enter the profile in the text boxes. If the option **Auto number** is unchecked in the *Profiles tab*, specify a unique sequence type number in the **Sequence type** text box. If the **Auto number** option is checked, the **Sequence type** text box is not editable and the software will automatically assign a new ID to the new profile when the profile is added to the database.

When the profiles contain the correct information, press **<OK>** in the *Update MLST alleles/profiles* dialog box to add all new profiles in the list to the database.

In a similar fashion as for alleles, profiles can be manually added, updated or deleted from the database (see 5.5.2).

Figure 5.10: Adding an allelic profile to the custom MLST database.

5.5.4 Editing clonal complexes

ST	CC	Status
1	ST1-complex	NEW

Figure 5.11: Clonal complex information in the custom MLST database.

In the *Clonal Complexes* tab, clonal complex information can be added to the database.

When pressing the <**Screen selection**> button, the software compares the clonal complex and sequence type combinations of the current entry selection to the information that is already present in the custom database. All newly detected clonal complex and sequence type combinations are added to the list in the *Clonal Complexes* tab.

The **ST** column displays the name of the sequence type, the **Clonal complex** column holds the clonal complex information, and the last column displays the **Status** information.

Double-click on a clonal complex and sequence type combination in the list if you want to edit the information. This pops up the *Edit clonal complex* dialog box.

A new clonal complex can also be added manually by pressing the <**Add**> button. This calls the *Edit clonal complex* dialog box (see Figure 5.12).

Pressing <**OK**> adds the new sequence type and clonal complex combination to the database.

When the clonal complexes contain the correct information, press <**OK**> to add all new clonal complexes in the list to the database.

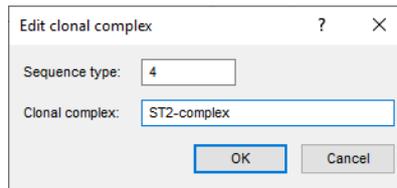


Figure 5.12: Add clonal complex to the custom MLST database.

In a similar fashion as for alleles and profiles, clonal complexes can be manually added, updated or deleted from the database (see 5.5.2).

5.6 Exporting alleles and profiles

Locally stored allele and profile information can be exported as FASTA-formatted text files from the database with the command **MLST > Export alleles and profiles**. This feature can be particularly useful when moving from an MLST setup with the allele and profile information stored in the database to a setup in which the information is made available through files, stored on a network drive or located on the internet.

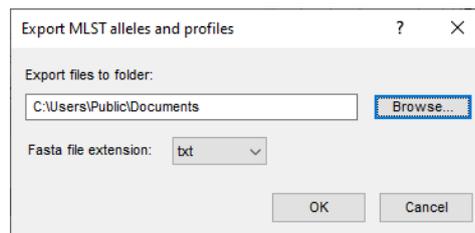


Figure 5.13: The *Export MLST alleles and profiles* dialog box: Export locally stored allele and profile information.

Browse for an export folder and select a **FASTA file extension** from the drop down list.

Pressing <**OK**> exports the MLST allele and profile information. For each sequence type, BIONUMERICS creates a separate FASTA file, containing the sequences that are locally stored in the database. The allelic profiles, sequence types and clonal complex information are stored in the tab-delimited `Profiles.txt` file.

5.7 MLST search

The MLST search tool is designed for searching entries with certain allelic profiles, corresponding to combinations of character states in the **MLST** character data set. Selecting **MLST > MLST search** calls the *MLST search* dialog box (see Figure 5.14).

The *MLST search* dialog box lists all loci that are defined in the MLST scheme. The user can enter the selection criteria for each of the loci in the *Alleles panel*. The criteria for the different loci are combined with AND. For each locus, comma separated values can be specified, which are then combined with OR.

The selection criteria can be cleared with the <**Clear search**> button. In the right panel, a **Maximum number of mismatches** can be specified. With the **Negative search** option checked, all

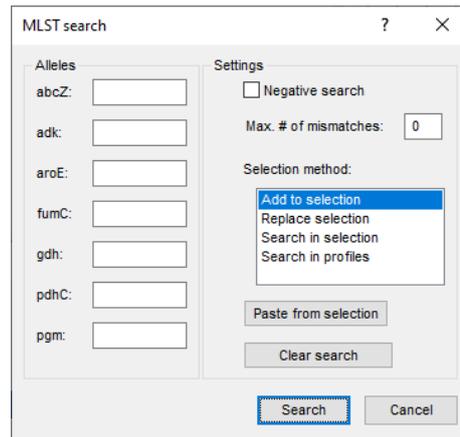


Figure 5.14: The *MLST search* dialog box.

entries that do not match the specified criteria will be selected. One can specify to add the entries to the current selection (**Add to selection**), replace an existing selection of entries by the found entries (**Replace selection**), or to search within an existing selection of entries (**Search in selection**). With **Search in profiles**, the profiles in the database are searched. Search criteria can be pasted from the character states of the currently selected entries using the **Paste from selection** button.

When pressing the **Search** button all entries in the database that match the specified criteria are selected and are brought to the top of the *Database entries* panel. In case **Search in profiles** was selected, a window will pop up, listing all sequence types that fulfilled the search criteria.

Chapter 6

MLST data analysis

Some useful features in the context of MLST data analysis will be highlighted in this chapter. More detailed information about the analysis possibilities in the software can be found in the BIONUMERICS manual.

6.1 Selections in BIONUMERICS

- 1.1 Select a single entry in the *Database entries* panel by holding the **Ctrl**-key and left-clicking on the entry. Alternatively, use the **space bar** to select a highlighted entry or click the ballot box next to the entry.

Selected entries are marked by a checked ballot box (☑) and can be unselected in the same way.

- 1.2 In order to select a group of entries, hold the **Shift**-key and click on another entry.

A group of entries can be unselected the same way.

- 1.3 All entries can be selected at once with **Edit > Select all (Ctrl+A)**.

- 1.4 Clear all selected entries with **Database > Entries > Unselect all entries (all levels) (F4)**.

6.2 The Comparison window

- 2.1 Make a selection in the *Database entries* panel (see 6.1).

- 2.2 Highlight the *Comparisons* panel in the *Main* window and select **Edit > Create new object... (+)** to create a new comparison for the selected entries.

A *Comparison* window is created, with the selected database entries.

- 2.3 You can drag the vertical separator lines between the panels to the left or to the right, in order to divide the space among the panels optimally.

- 2.4 Click on the  next to the experiment name **MLST** in the *Experiments* panel and select **Characters > Show values (☐)** to display the allele numbers in the *Experiment data* panel (see Figure 6.1).

- 2.5 In the *Comparison* window, groups can be created based on a selection (**Groups > Create new group from selection (📌, Ctrl+G)**) or based on the content of an information field: right-click in the header of a field (e.g. field **MLST CC**) and select **Create groups from database field** from the menu.

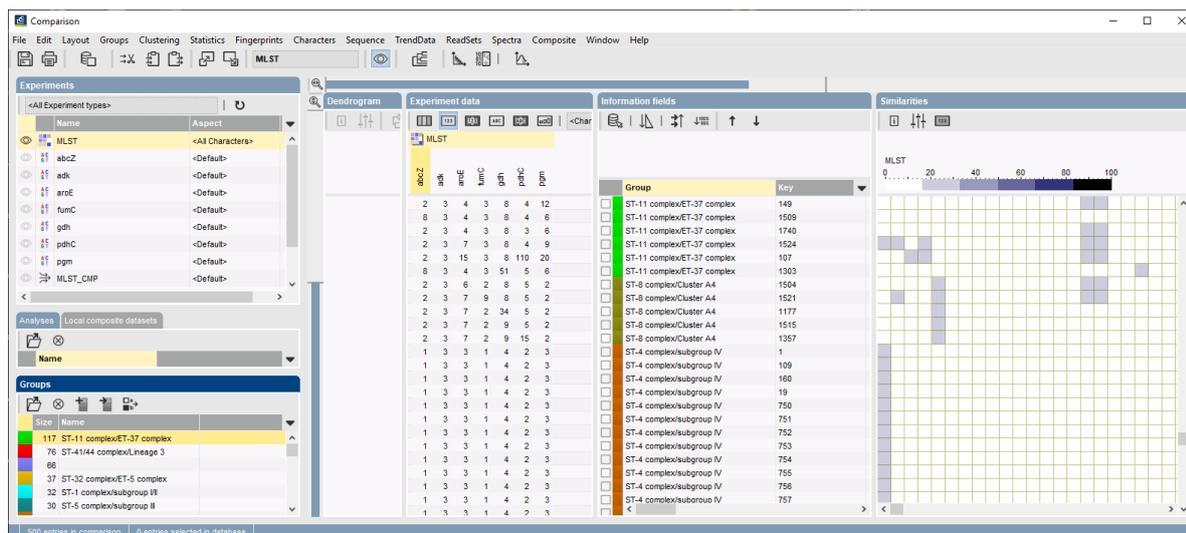


Figure 6.1: The *Comparison* window: groups have been defined based on the **MLST** **CC** information field.

- 2.6 Click on the  next to a sequence experiment name in the *Experiments* panel to view the sequences in the *Experiment data* panel.
- 2.7 Click on the  next to the experiment name **MLST_CMP** in the *Experiments* panel to view the mutation list of the concatenated MLST sequences in the *Experiment data* panel.

6.3 Cluster analysis

- 3.1 In the *Experiments* panel of the *Comparison* window, make sure the **MLST** experiment is selected.
- 3.2 Select **Clustering** > **Calculate** > **Cluster analysis (similarity matrix)...** to display the *Similarity coefficient wizard* page.

Due to the arbitrariness of the allele numbers, the only suitable similarity coefficients for clustering MLST data are categorical coefficients. A categorical coefficient compares the allele numbers to see if they are the same or different but does not quantify the difference.

- 3.3 Select e.g. the **Categorical (values)** coefficient from the list and press <**Next**>.

- 3.4 In the *Cluster analysis* wizard page, choose a clustering method (e.g. **UPGMA**) and press <**Finish**>.

When finished, the dendrogram and the similarity matrix are shown in the *Comparison* window (see Figure 6.2 for an example).

A minimum spanning tree in BIONUMERICS is calculated in the *Advanced cluster analysis* window. This window can be launched from the *Comparison* window:

- 3.5 Select **Clustering** > **Calculate** > **Advanced cluster analysis...** or press the  button and select **Advanced cluster analysis** to launch the *Create network wizard*.

The predefined template **MST for categorical data** uses the categorical coefficient for the calculation of the similarity matrix, and will calculate a standard minimum spanning tree with single and

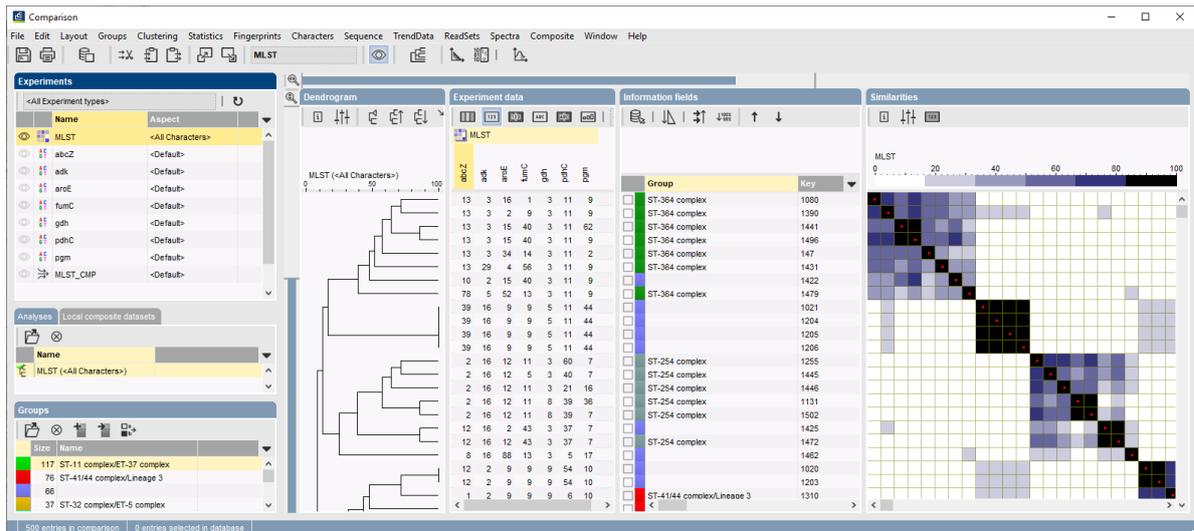


Figure 6.2: The *Comparison* window with a UPGMA tree displayed.

double locus variance priority rules.

3.6 Specify an analysis name (for example **MLST1**), make sure **MLST(<All Characters>)** is selected, select **MST for categorical data**, and press **<Next>**.

The *Advanced cluster analysis* window pops up (see Figure 6.3 for an example).

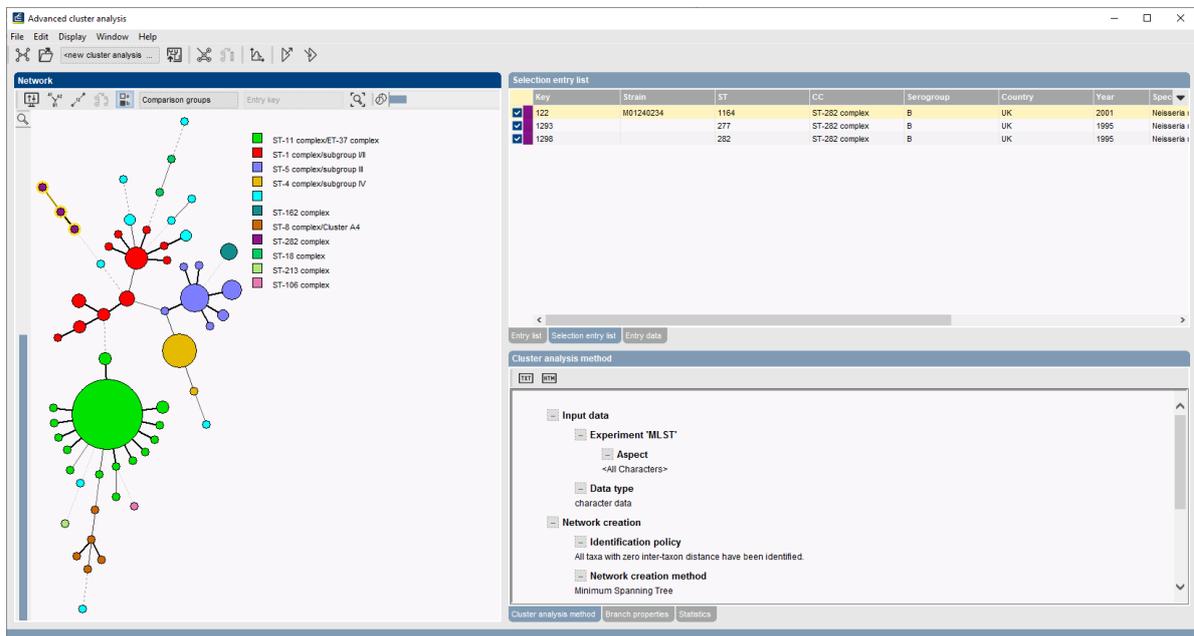


Figure 6.3: The *Advanced cluster analysis* window.

The *Network panel* in the *Advanced cluster analysis* window displays the minimum spanning tree, the upper right panel (*Entry list*) displays the entries that are present in the tree. The *Cluster analysis method panel* displays the settings used, in this example the priority rules that result in the displayed network. The colors of the comparison groups (if defined) are automatically shown as node colors.

3.7 To change the display settings of the network (node/branch labels, node/branch colors, etc.), press the  button or choose **Display > Display settings**.

- 3.8 Select **File** > **Exit** to close the *Advanced cluster analysis* window and select **File** > **Save** (📁, **Ctrl+S**) to save the comparison.

All calculations done on the data is stored along with the comparison. This includes a similarity matrix for the experiment type on which a dendrogram was last calculated, any dendrogram that has been created, any statistical analysis, comparison groups, etc..

- 3.9 Enter a name, e.g. "MyComp" and press <**OK**>.

- 3.10 Close the comparison with **File** > **Exit**. The comparison **MyComp** is now listed in the *Comparisons* panel of the *Main* window.

To continue working on an existing comparison, it can be opened again:

- 3.11 Highlight the comparison in the *Comparisons* panel and select **Edit** > **Open highlighted object...** (📁, **Enter**). Alternatively, just double-click on the comparison name.

