BioNumerics Tutorial:

Extracting subsequences 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 screen the whole genome sequences of some *Staphylococcus aureus* samples for the *mecA* sequence. The different steps are illustrated using the whole genome demonstration database of *Staphylococcus aureus*. This database is available for download on our website (see 2) and contains 97 publicly available sequence read sets of *Staphylococcus aureus* with already calculated de novo assemblies.

2 Preparing the database

2.1 Introduction to the demonstration database

The whole genome demonstration database of *Staphylococcus aureus* 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).

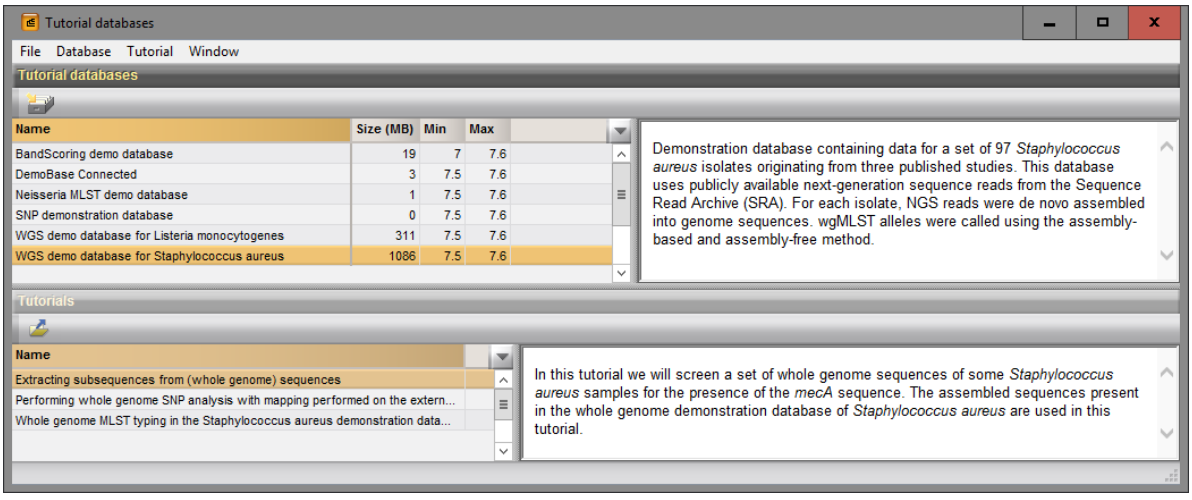

Figure 1: The *Tutorial databases* window, used to download the demonstration database.

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

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

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

The **WGS demo database for Staphylococcus aureus** appears in the *BioNumerics Startup* window.

5. Double-click the **WGS demo database for Staphylococcus aureus** 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 whole genome demo database for Staphylococcus aureus is also available on our website. This backup can be restored to a functional database in BioNumerics.

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

In contrast to other browsers, some versions of Internet Explorer rename the `wgMLST_SAUR.bnbk` database backup file into `wgMLST_SAUR.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. "Whole genome Staphylococcus aureus demobase".

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

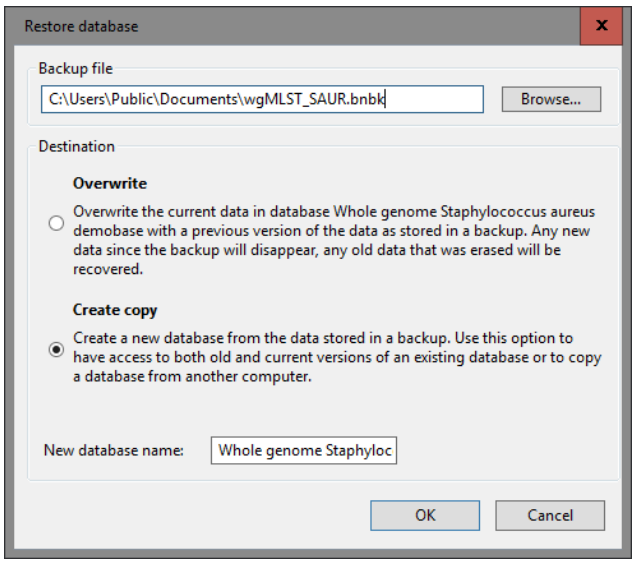

Figure 2: Restoring the whole genome demonstration database from the BioNumerics backup file `wg_SAUR.bnbk`.

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

3 About the demonstration database

The whole genome demonstration database of *Staphylococcus aureus* contains links to sequence read set data on NCBI's sequence read archive (SRA) for 97 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 3).

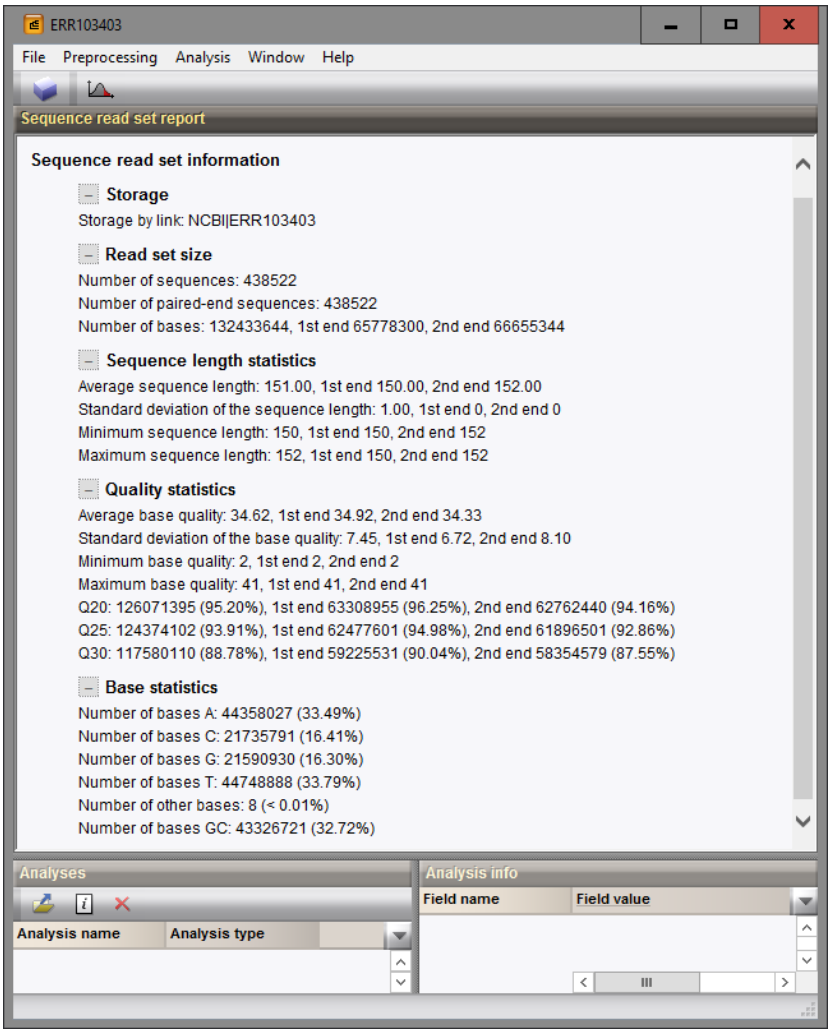

Figure 3: The sequence read set experiment card for an entry.

2. Close the *Sequence read set experiment* 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 4).

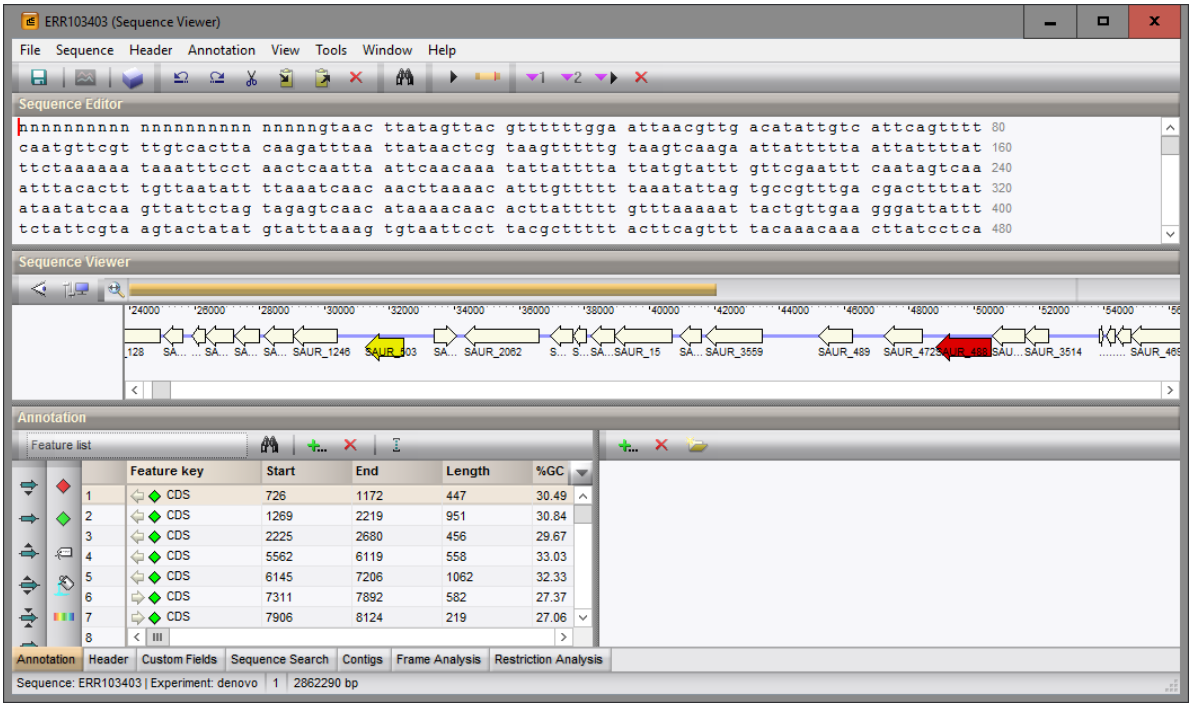

Figure 4: The *Sequence editor* window.

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 experiment. We will search for the *mecA* sequence.

5. Close the *Sequence editor* window.

Additional information, stored in entry info fields (Organism name, Instrument, Study accession, 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 Staphylococcus aureus 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 5).

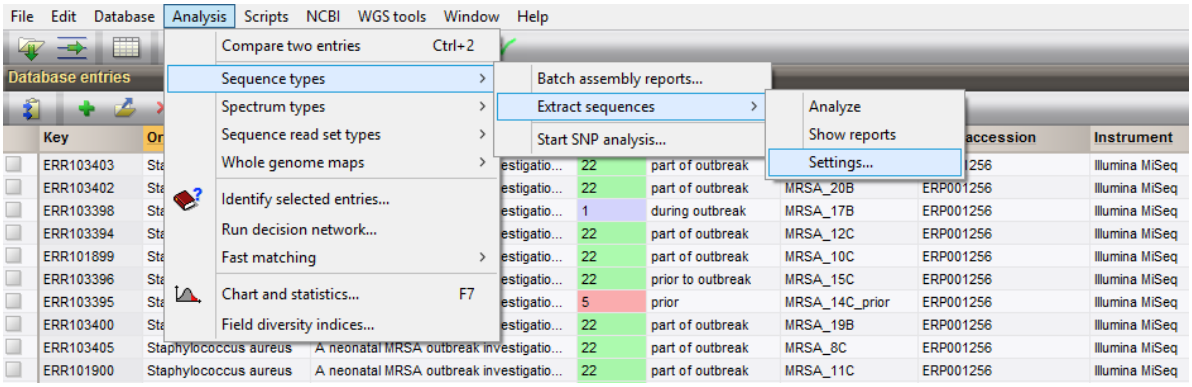

Figure 5: Additional menu items installed by the *Sequence extraction plugin*.

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 6). This dialog is called via ***File*** > ***Process...*** ( ).

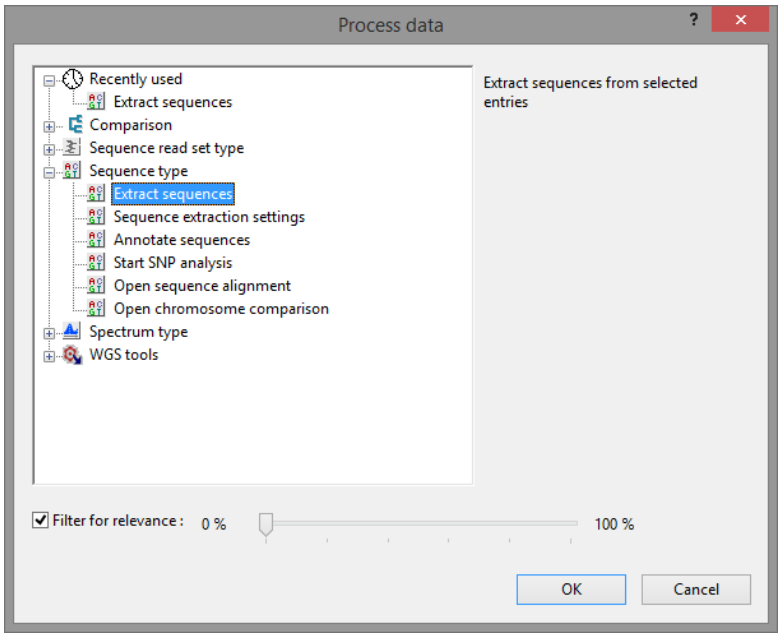

Figure 6: The *Process data* dialog box, displaying the two items (***Extract sequences*** and ***Sequence extraction settings***) that are injected by the *Sequence extraction plugin*.

5 Extracting subsequences

5.1 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 (see 5.2), and specify the sequence extraction

settings (see 5.3).

5.2 Provide query sequence

A FASTA formatted text file can be found on our website, containing the *mecA* sequence extracted from the reference sequence with accession number **BX571856.1** (see Figure 7). The example file can be found on the download page on our website (`http://www.applied-maths.com/download/sample-data`, "mecA reference sequence").

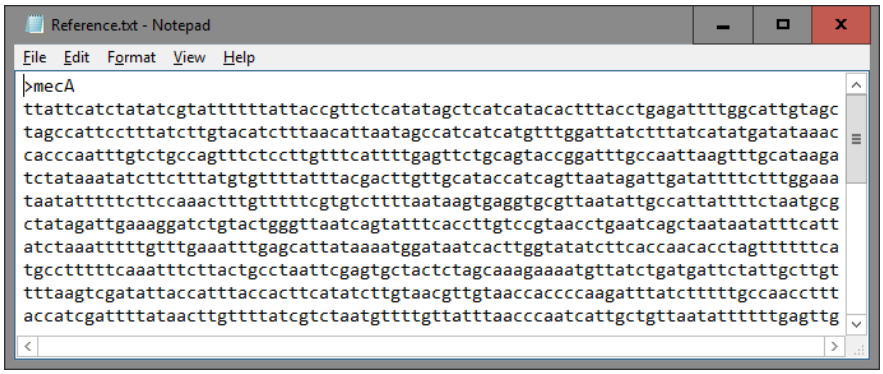

Figure 7: mecA reference sequence.

In order to use this sequence as query sequence we first need to import this sequence 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 `Reference.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 gene (see Figure 8). We will use this name for our sequence experiment in our database.

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 9).

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

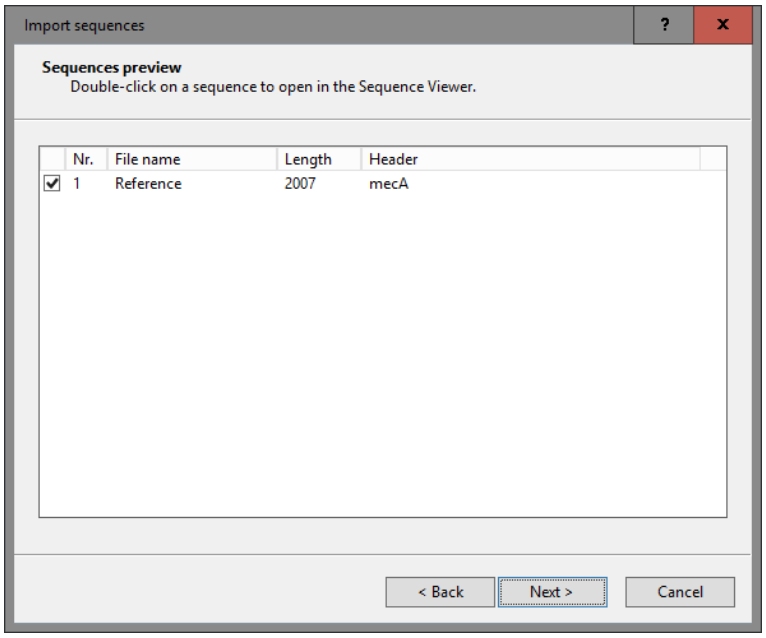

Figure 8: Preview.

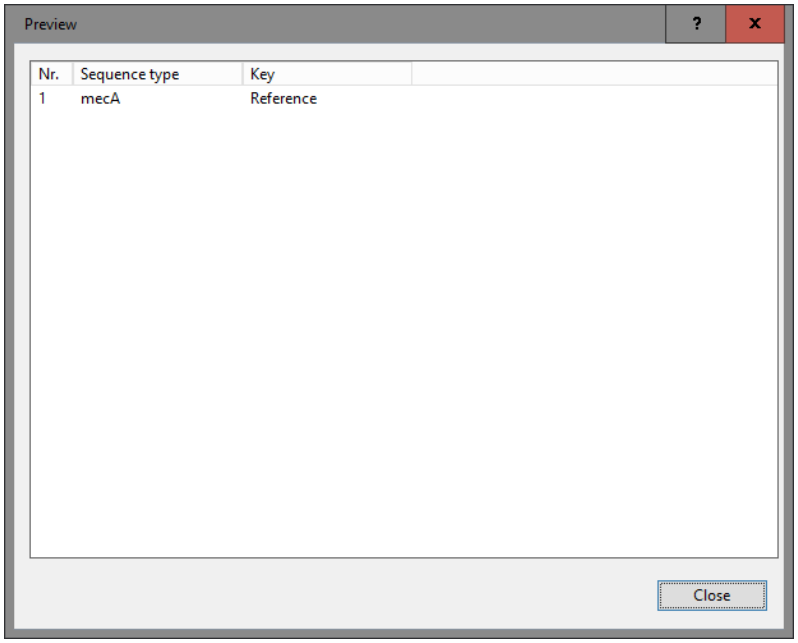

Figure 9: Preview.

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

12. Highlight the newly created template (see Figure 10) and press <***Next***>.

13. Press <***Yes***> twice to confirm the creation of the ***mecA*** sequence experiment in the database (see Figure 11).

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

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

An entry with key ***Reference*** is created in the database and the sequence from the text file is linked to the

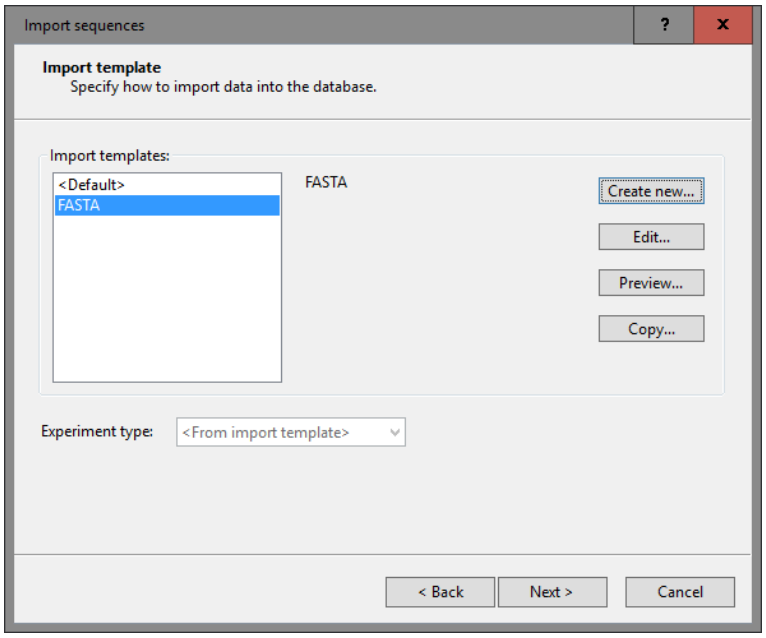

Figure 10: Import template.

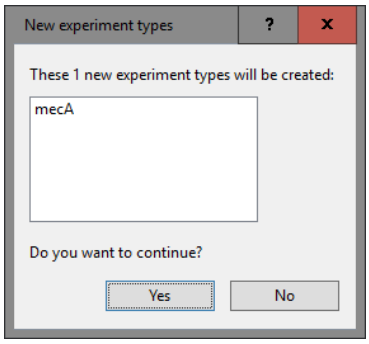

Figure 11: Add new sequence type to the database.

mecA sequence experiment.

5.3 Specify sequence extraction settings

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

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.

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

17. Select ***denovo*** as ***Origin experiment type*** (see Figure 12). This is the sequence experiment, containing the whole genome sequences, that will be screened and from which a subsequence will be copied from.

18. Check the ***mecA*** experiment as ***Destination experiment type*** (see Figure 12).

19. 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

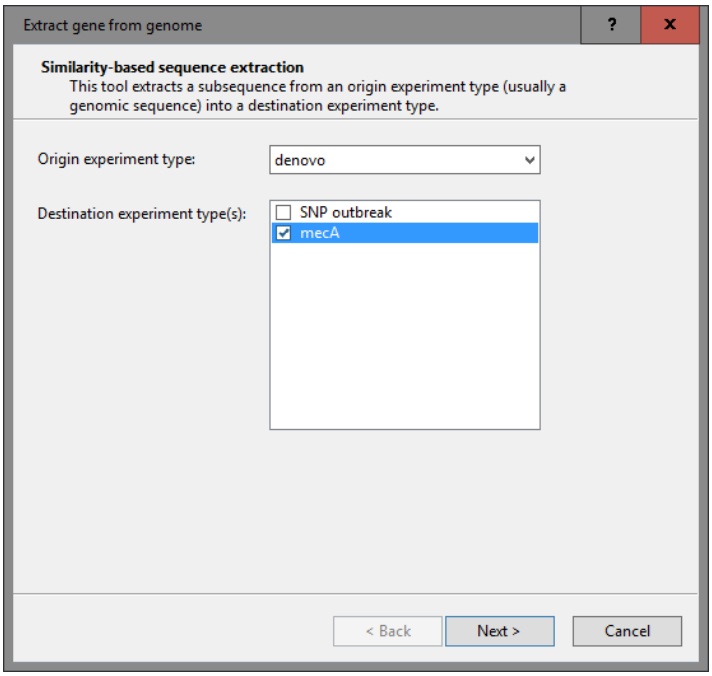

Figure 12: Specify the origin and destination sequence types.

denovo) for. In our demonstration database, entry with key ***Reference*** contains the query sequence, stored in the ***Destination experiment type***.

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

21. Scroll down the list, highlight ***Reference*** and press <***OK***>. Alternatively, start typing the text ***Reference*** in the ***Search for*** text box, highlight the reference entry 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).

22. For this exercise, make sure the ***Reference*** 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 14).

Default report settings will be applied when running a report (see 5.4), but can be modified by highlighting ***Reports*** in the tree and pressing <***Edit***>.

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

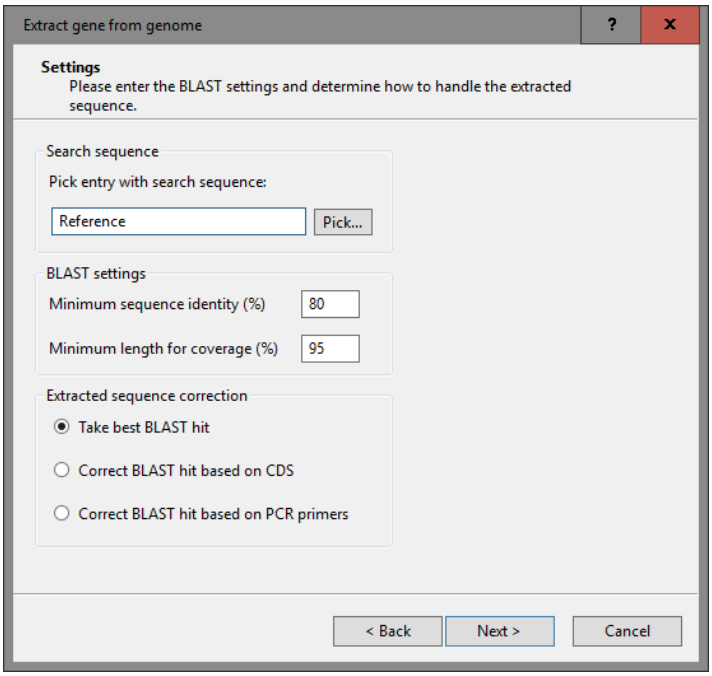

Figure 13: Sequence extraction settings.

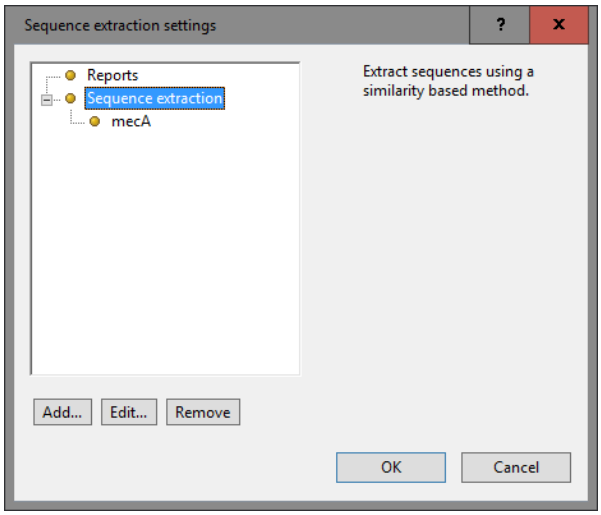

Figure 14: Sequence extraction settings.

5.4 Sequence extraction analysis

Now that we have specified the sequence extraction settings (see 5.3), we can now start the actual sequence extraction process.

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

The status bar, displayed at the bottom of the *Main* window will indicate that **97** entries are selected.

25. 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.

26. 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 15).

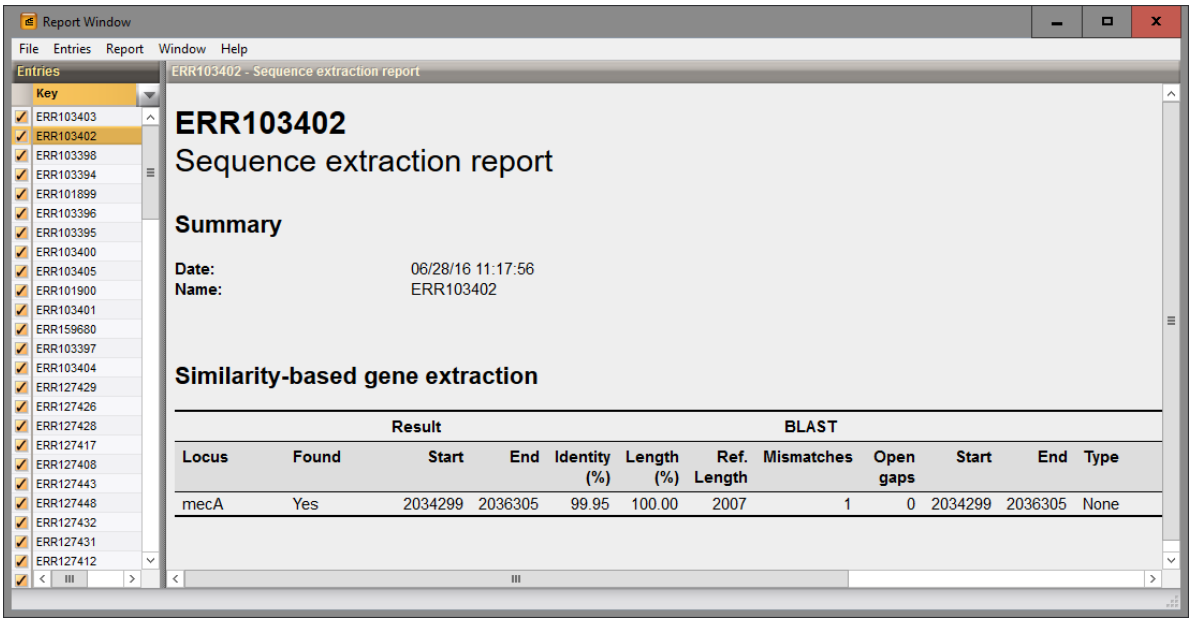

Figure 15: Summary of the sequence extraction results.

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 (here: only one), 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.

27. Close the *Report* window.

The extracted sequences - if found - are stored in the ***mecA*** sequence experiment type.

28. Clicking on a green colored dot in the *Experiment presence* panel for a ***mecA*** experiment will open the *Sequence editor* window, containing the extracted sequence (see Figure 16).

29. Close the *Sequence editor* window.

In the tutorial ”MLST analysis starting from whole genome sequences” available on our website, the *Sequence extraction plugin* is used to extract the seven MLST housekeeping genes from a set of whole genome sequences of *Listeria monocytogenes*. Follow-up analysis is illustrated in combination with the *MLST online plugin*.

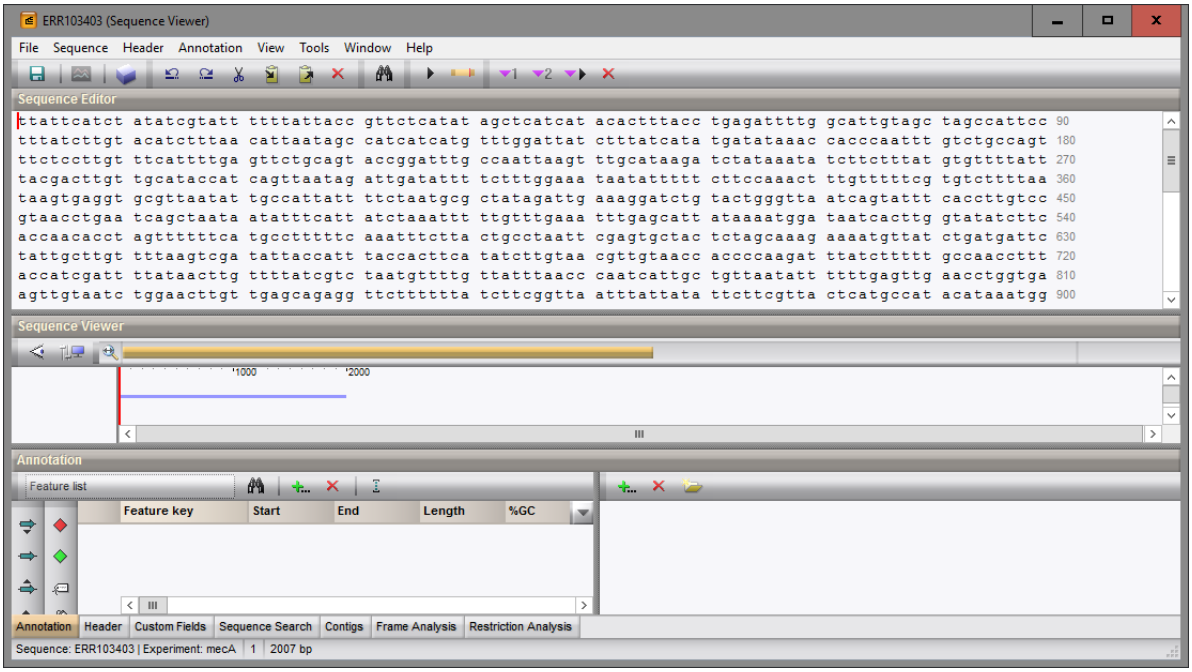

Figure 16: The *Sequence editor* window.