BIONUMERICS Tutorial:

Mycobacterium tuberculosis complex functional genotyping: predicting phenotypic traits from whole genome sequences

1 Aim

In this tutorial we will screen whole genome sequences of *Mycobacterium tuberculosis* complex samples for phenotypic traits such as spoligotype, lineage and antibiotic resistance using the *Mycobacterium tuberculosis complex functional genotyping plugin*. The plugin also allows you to perform species confirmation.

The different steps are illustrated using the whole genome demonstration database of the *Mycobacterium tuberculosis* complex. This database is available for download on our website (see 2) and contains 29 publicly available sequence read sets of the *Mycobacterium tuberculosis* complex with already calculated de novo assemblies.

2 Preparing the database

2.1 Introduction to the demonstration database

We provide a **WGS demo database** for the *Mycobacterium tuberculosis* complex containing sequence read set data links for 29 samples, calculated de novo assemblies and wgMLST results (allele calls and quality information).

The wgMLST workflow and results will not be discussed in this tutorial.

The **WGS demo database** for the *Mycobacterium tuberculosis* complex 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).

Installation of the *Mycobacterium tuberculosis complex functional genotyping plugin* is only possible when no spaces are present in the BIONUMERICS home directory and in the name of the database. Before downloading or restoring the **WGS demo database** for the *Mycobacterium tuberculosis* complex, please check if your BIONUMERICS home directory does not contain any spaces:

1. Click the  button, located in the toolbar in the *BIONUMERICS Startup* window and select ***Change home directory...*** to call the *Home directory* dialog box.

2. In case the currently specified home directory contains spaces, update the path to a path containing no spaces and close the *Home directory* dialog box.

2.2 Option 1: Download demo database from the Startup Screen

3. Click the  button, located in the toolbar in the *BIONUMERICS Startup* window.

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

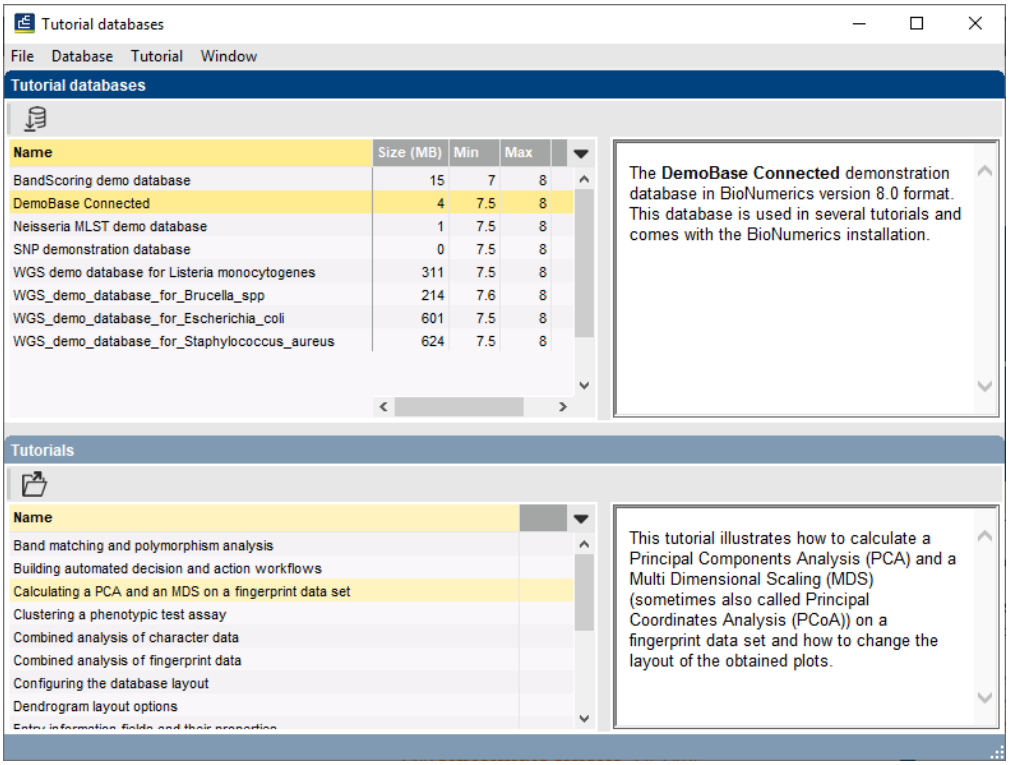

Figure 1: The *Tutorial databases* window, used to download the demonstration database.

4. Select **WGS_demo_database_for_MTBC** from the list and select ***Database*** > ***Download*** ( ).

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

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

The **WGS_demo_database_for_MTBC** appears in the *BIONUMERICS Startup* window.

7. Double-click the **WGS_demo_database_for_MTBC** 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 demo database for the *Mycobacterium tuberculosis* complex is also available on our website. This backup can be restored to a functional database in

BIONUMERICS.

8. Download the file `WGS_MTBC.bnbk` file from https://www.applied-maths.com/download/sample-data, under 'WGS_demo_database_for_MTBC'.

In contrast to other browsers, some versions of Internet Explorer rename the `WGS_MTBC.bnbk` database backup file into `WGS_MTBC.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.

9. In the *BIONUMERICS Startup* window, press the  button. From the menu that appears, select ***Restore database...***.
10. Browse for the downloaded file and select ***Create copy***. Note that, if ***Overwrite*** is selected, an existing database will be overwritten.
11. Specify a new name for this demonstration database, e.g. "WGS_MTBC_demobase".
12. Click <***OK***> to start restoring the database from the backup file.
13. Once the process is complete, click <***Yes***> to open the database.

The *Main* window is displayed (see Figure 2).

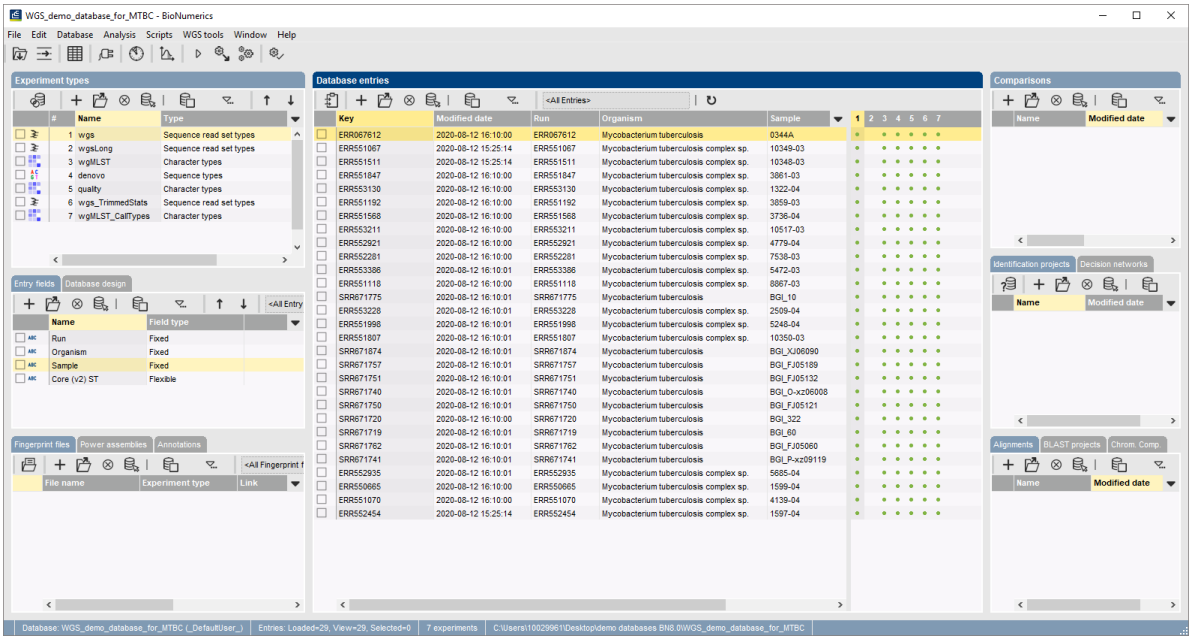

Figure 2: The *MTBC* demonstration database: the *Main* window.

3 About the demonstration database

The WGS demo database contains links to sequence read set data on NCBI's sequence read archive (SRA) for 29 publicly available sequencing runs. Additional information (in entry info fields Sample and Organism) was collected from NCBI and added to the demonstration database.

Seven experiments are present in the demo database and are listed in the *Experiment types* panel (see Figure 3).

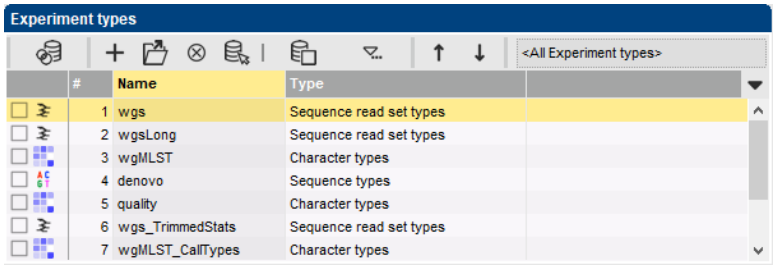

Figure 3: The *Experiment types* panel in the *Main* window.

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.

3. Click on the green colored dot for one of the entries in the fourth column in the *Experiment presence* panel. Column 4 corresponds to the fourth 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. Close the *Sequence editor* window.

The sequence read set experiment type **wgs_TrimmedStats** contains some data statistics about the reads retained after trimming, used for the de novo assembly.

The sequence read set experiment type **wgsLong** contains the links to long read sequence read data (typically PacBio or MinION datasets). In this demo database, no links are defined for this experiment.

The other three experiments contain data related to the wgMLST analysis performed on the samples:

- 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.

- Character experiment type **quality** contains quality statistics for the raw data, the de novo assembly and the different allele identification algorithms.

- Character experiment type **wgMLST_CallTypes**: contains details on the call types.

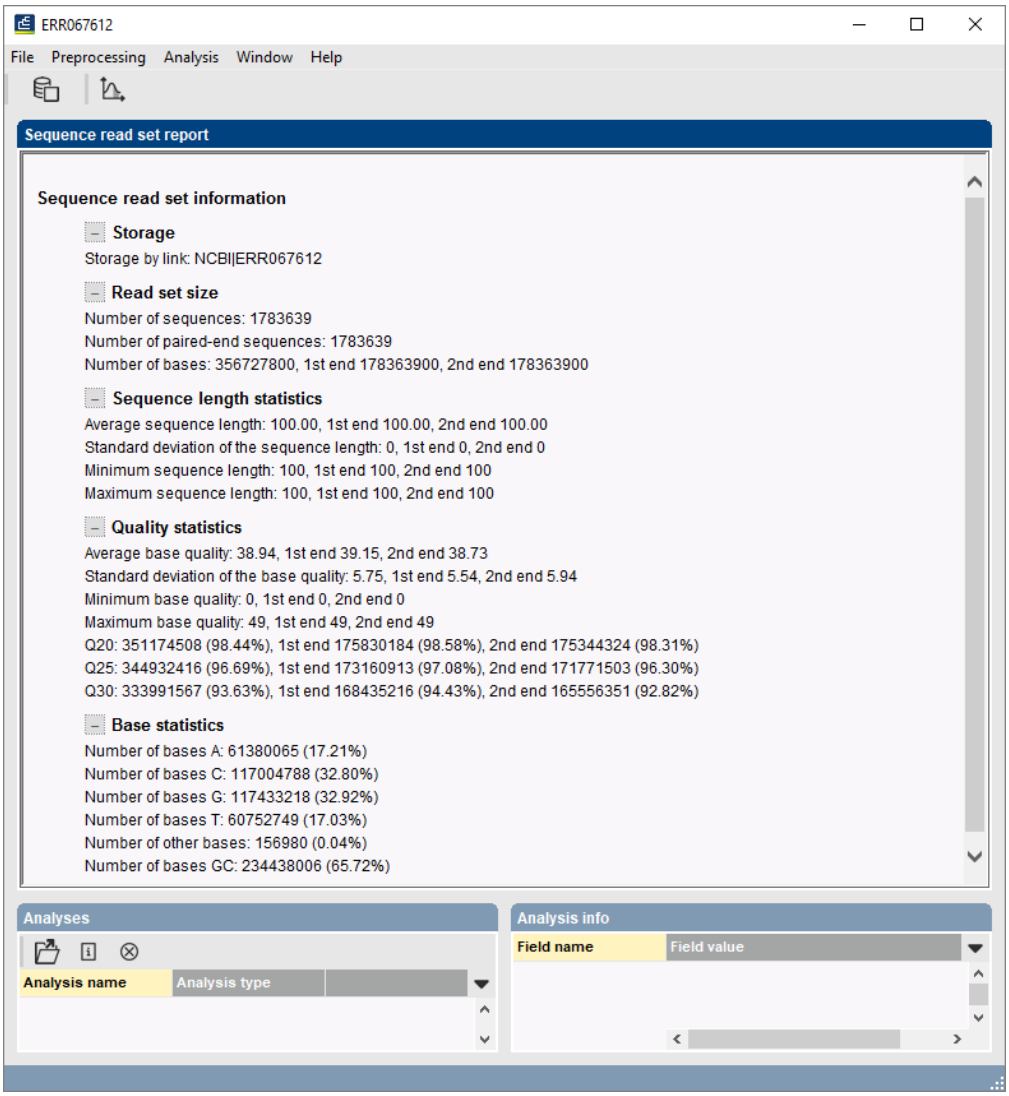

Figure 4: The sequence read set experiment card for an entry.

4 Installing the *MTBC* functional genotyping plugin

To store the results of your genotyping jobs, you will have to create some new information fields and character/sequence experiments.

1. Make sure the *Database entries* panel is the active panel in the *Main* window and select ***Edit*** > ***Information fields*** > ***Add information field...***.

2. Enter a name, e.g. ***Species confirmation*** and choose ***Space*** next to ***Optimize storage for*** (see Figure 6). The latter is important because otherwise results can get truncated if they contain too much characters. All other settings can be left default.

3. Repeat previous step for following information fields. Make sure you optimize the storage for ***Space*** each time:

- Unknown genotype
- Resistance summary
- Spoligotype (octal)

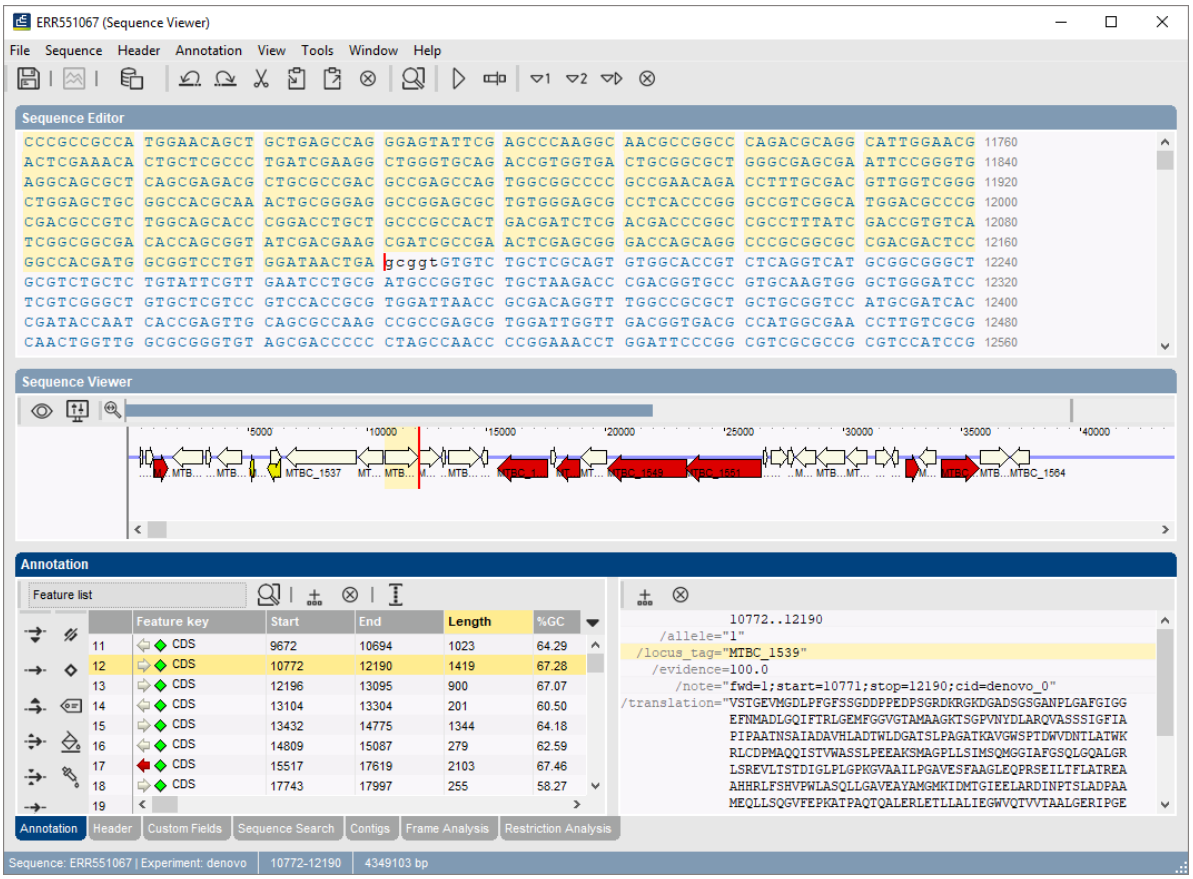

Figure 5: The *Sequence editor* window.

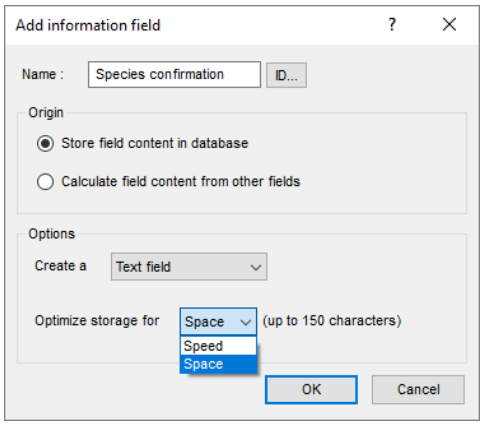

Figure 6: Adding a new information field with the ***Space*** option.

- Lineage number
- Lineage name

Proceed as follows to install the *MTBC genotyping plugin*:

4. Call the *Plugins* dialog box from the *Main* window with ***File*** > ***Install / remove plugins...*** ( ).
5. Select *MTBC functional genotyping* from the list in the *Applications tab* and press the <***Activate***> button.
6. Confirm the installation of the plugin.

7. In the *General tab* of the wizard (see Figure 7) choose ***wgs*** as ***WGS experiment type*** and the ***Info fields*** that will appear in the report. The ***Key*** and the ***Resistance/Lineage*** results are default shown in the report. Check for example the ***Species confirmation*** and ***Spoligotype (octal)*** information field or any other field that was added. Click <***Next***>.

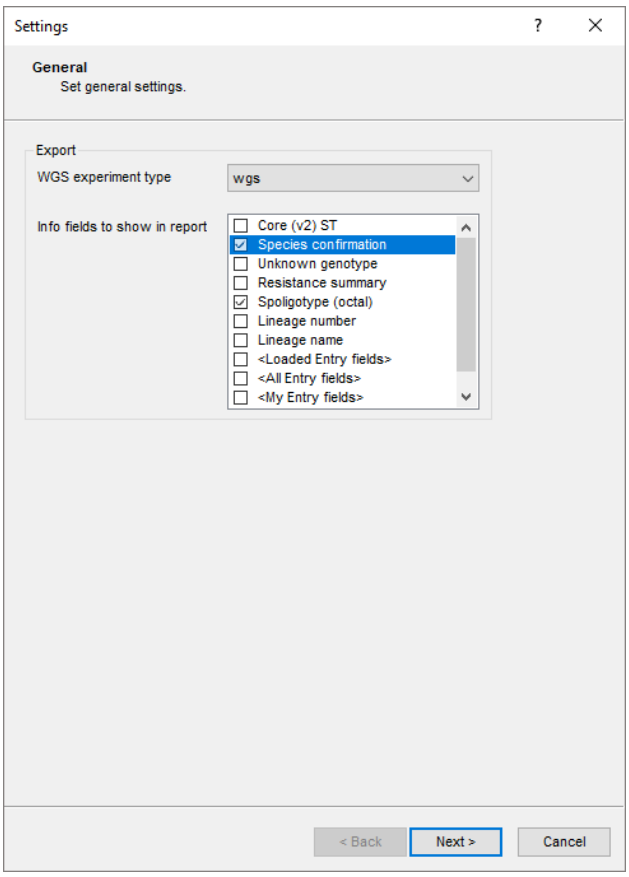

Figure 7: The *General tab*.

8. In the *Species tab* of the wizard choose the ***Species field*** for the storage of the species confirmation results, change the detection parameters if wanted and click <***Next***> (see Figure 8).

9. In the *Lineage tab* choose the fields for the storage of the ***Lineage number*** and ***Lineage name***, change the detection parameters if wanted and click <***Next***> (see Figure 9).

10. In the *Spoligotyping tab* select ***Create New*** as ***Spoligo presence/absence experiment type***. Choose the field for the storage of the ***Octal spoligotype***. Modify the detection parameters if wanted and click <***Next***> (see Figure 10).

11. If ***Create new*** was selected in the previous step, a dialog box appears prompting for the ***Spoligotype absence/present*** experiment type name. Enter e.g. **Spoligotype** and click <***OK***>.

12. In the *Resistance tab* (see Figure 11), select the ***Resistance database***. Currently, only one database is available: ***Curated database (version 6.0)***.

13. Select ***Create new*** next to the three ***Resistance results*** experiment types (see Figure 11).

14. Choose the ***Resistance summary*** information field for the storage of the antibiotics for which known resistance-related mutations (or indels) were found.

15. Choose the ***Unknown genotype*** information field for the storage of the antibiotics for which unknown mutations or indels (which are not included in the resistance database) were found in known resistance-related genes.

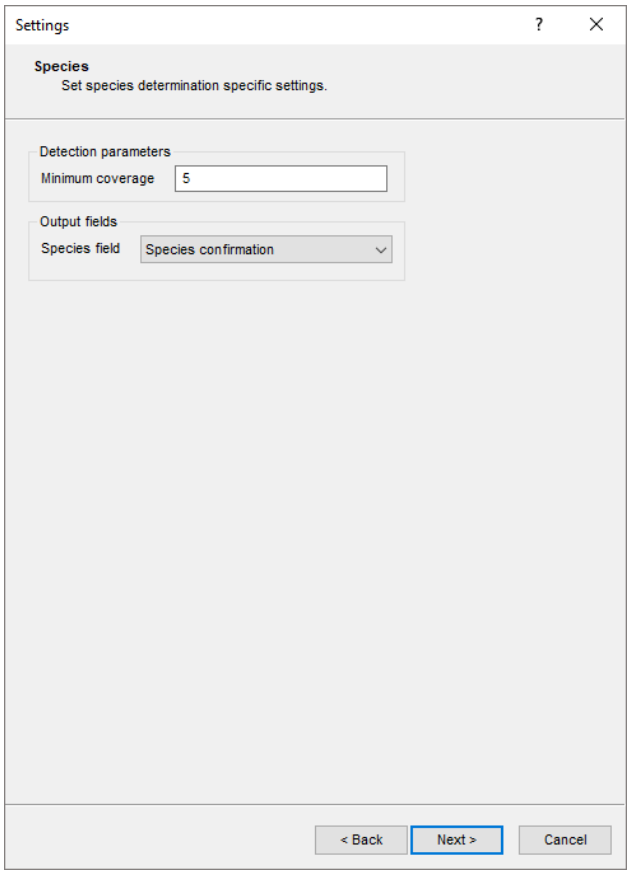

Figure 8: The *Species tab*.

16. Check ***Detect genomic variants*** if you want to store all the mutations in resistance-related genes in an experiment type. Select ***Create new*** or select an existing experiment from the ***Nucleotide variants*** and ***Amino acid variants*** experiment type lists.

17. Click <***Finish***>.

18. If ***Create new*** was selected in the previous step, a dialog box appears prompting for the experiment type name(s). Enter the name(s) and click <***OK***>.

When the *MTBC functional genotyping* installation is complete, you will be prompted to restart the database. The *Plugins* dialog box can be closed by pressing the <***Exit***> button and the database via ***File*** > ***Exit***.

Open the database again from the *BIONUMERICS Startup* window. A ***MTBC*** menu item is now available in the *Main* window (see Figure 12).

5 Screening of entries

5.1 Submitting jobs

MTBC jobs need to be submitted to the calculation engine. The Calculation engine option requires credits for running jobs on the Applied Maths cloud calculation engine. Credits are linked to credentials that you need to enter when installing the WGS tools plugin. In our demo database, the WGS tools plugin is installed but no credits are assigned to the demo project so no MTBC jobs

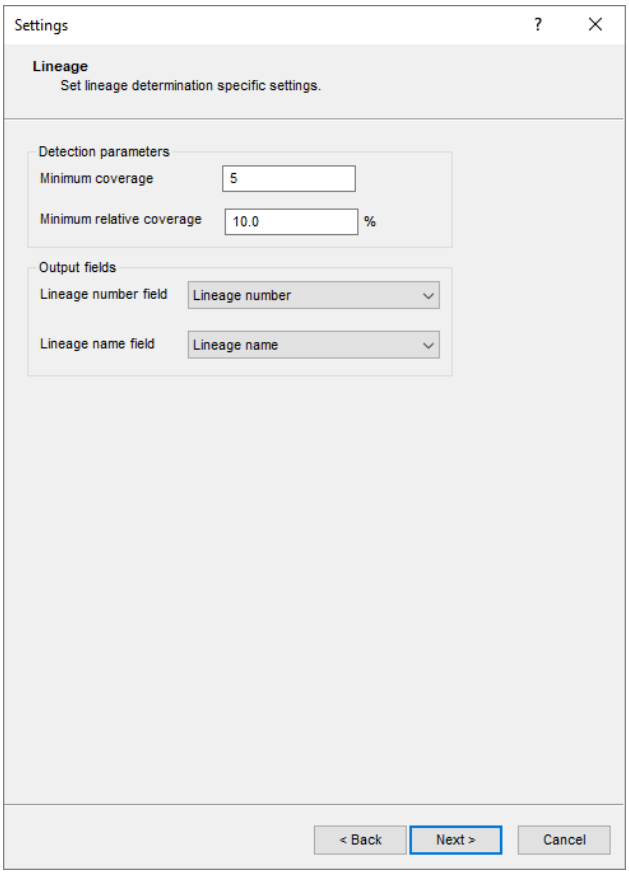

Figure 9: The *Lineage* tab.

can be performed on the external calculation engine. Please contact Applied Maths to obtain more information.

Once the *MTBC genotyping plugin* is installed and the settings have been specified, an MTBC genotyping job (spoligotyping, resistance/lineage prediction and species prediction) can be submitted.

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.

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

3. Select ***WGS tools*** > ***Submit jobs...*** ( ) to open the *Submit jobs* dialog box.

From the *Submit jobs* dialog box, one can define which algorithms need to be run on the selected samples and as such, define and launch the related jobs on the calculation engine.

4. Select ***Calculation Engine***.

5. Check the box next to ***MTBC genotyping*** (see Figure 13). Note that this option is only available after successful installation of the plugin and closing and reopening of the database after installation.

At this stage, you can also still change the settings by clicking the <***Settings***> button next to ***MTBC genotyping***.

6. Uncheck all other boxes if you do not want to perform any additional analyses (e.g. wgMLST).

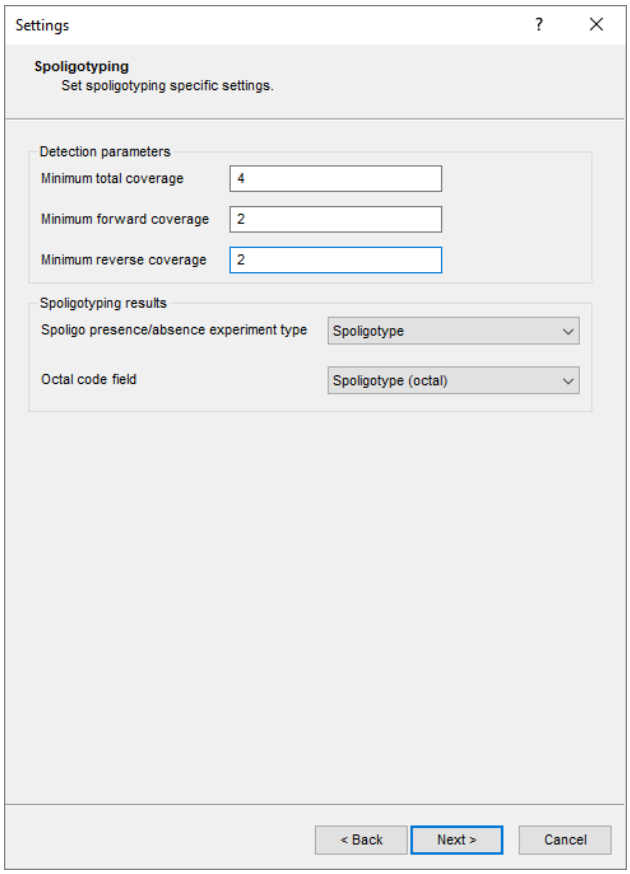

Figure 10: The *Spoligotyping* tab.

Jobs that already have been submitted and have been imported successfully, will not be re-launched for analysis, unless the check box in front of ***Re-submit already processed data*** in the ***Jobs*** part is checked.

By default, the *Job overview* window will be opened after submission of the jobs. However, this can be changed by unchecking the option ***Open jobs overview window***.

To analyze one sample with the MTBC genotyping tool, you need 2 credits. The number of credits required to run the selected jobs for the selected entries can be consulted at the bottom of the *Submit jobs* dialog box.

7. Click <***OK***> to launch the MTBC genotyping jobs and open the *Job overview* window for the calculation engine (see Figure 14).

In the *Job overview* window you can see the status of the submitted jobs. The *Job overview* window can be opened from the *Main* window with ***File*** > ***Jobs overview...*** ( ).

8. Finished jobs can be imported with a manual action (***Jobs*** > ***Get results*** ( )) or through an automatic update: select ***File*** > ***Settings***, check both options and specify an interval (e.g. 10 min).

The job results can also be imported starting from the entry selection in the *Main* window:

9. Make an entry selection in the *Database entries* panel and select ***WGS tools*** > ***Get results*** ( ).

The job log files are saved in the *Job log* panel of the *Entry* window. Double-click on an entry in the *Database entries* panel to open the *Entry* window and to consult this information.

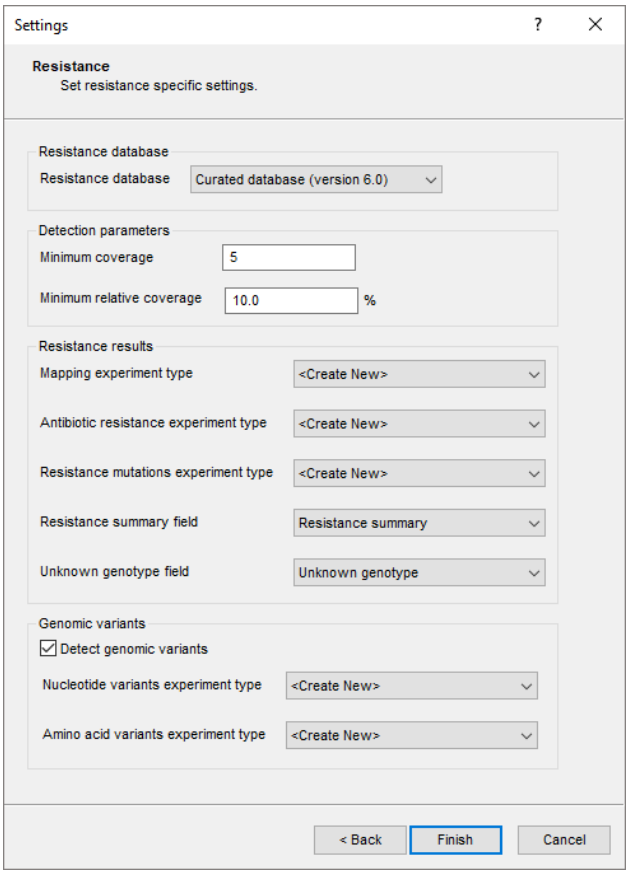

Figure 11: The *Resistance* tab.

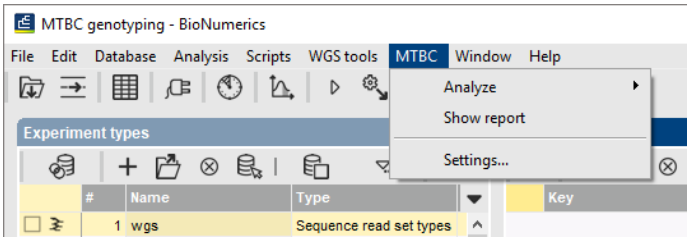

Figure 12: New menu items after installation of the plugin.

Once the results are imported, the corresponding jobs and their underlying data sets are automatically deleted from the calculation engine and as such, from the *Job overview* window.

5.2 Local screening

If you want to redo the lineage or resistance prediction of your samples with other settings, you do not need to submit a new job to the CE. This means that no additional credits will be charged:

1. Make new information fields to store the new lineage or resistance results (optional step).
2. Open the *Settings dialog box* with ***MTBC*** > ***Settings...*** in the *Main* window.
3. Select the new information fields and/or create new character experiments to store the new resistance results (optional step).

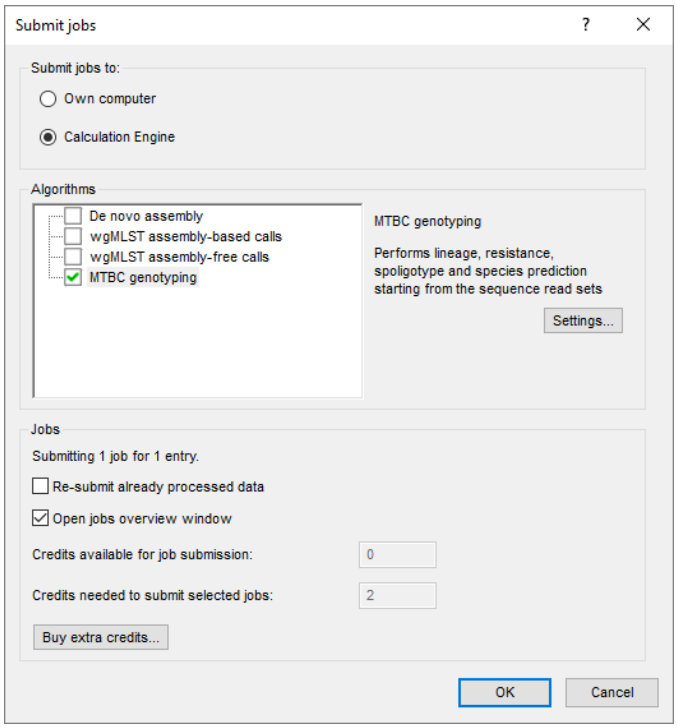

Figure 13: Submitting jobs to the calculation engine.

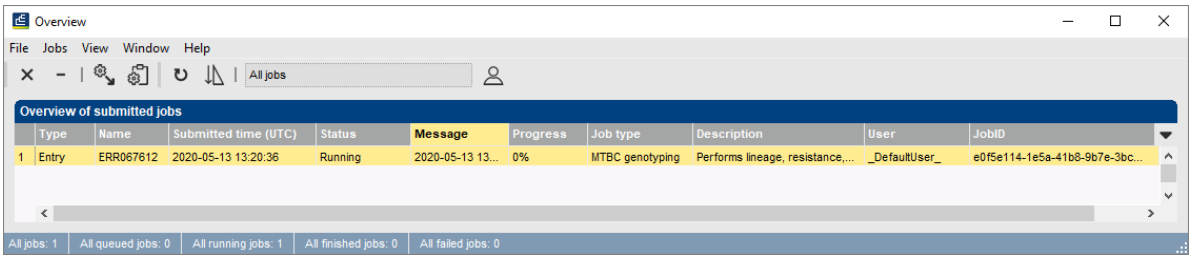

Figure 14: Job overview.

4. Select other settings.

5. Resistance and lineage prediction can be re-analyzed locally by clicking in the *Main* window on ***MTBC*** > ***Analyze*** > ***Resistance*** or ***MTBC*** > ***Analyze*** > ***Lineage*** respectively. If you want to redo both analysis, select ***MTBC*** > ***Analyze*** > ***All enabled***.

The options ***MTBC*** > ***Analyze*** > ***Resistance/lineage/all enabled*** can only be used when there is already a mapping present for the selected entries (this will only be the case if you have already submitted an MTBC genotyping job for this sample to the CE). Please note that only the mapping generated by the MTBC genotyping plugin can be used and not the standard mapping.

If you did not create/select new experiment types in the settings, results in the current experiment types will be overwritten when you re-analyze your samples.

If you click on ***MTBC*** > ***Analyze*** > ***Resistance***, and you also selected new, empty fields for lineage results, the lineage results obtained with the original settings will be used. If you also want to analyze the lineage with new settings, you have to click on ***MTBC*** > ***Analyze*** > ***Lineage*** or ***MTBC*** > ***Analyze*** > ***All enabled***.

5.3 Results

The ***Species***, ***Lineage***, ***Spoligotyping*** and ***Resistance*** results are written to the information fields in the *Database entries* panel (see Figure 15). Please note that the shown names of the information fields can be different in your case depending on whether you choose an alternative name during installation.

An additional information field called ***MDR_or_XDR*** is generated automatically during the first import of results (see Figure 15). MDR: multidrug resistance is defined as predicted resistance to INH and RMP. XDR: extensively drug resistance is defined as predicted resistance to INH, RMP, (CAP, KAN or AMK) and 1 of the FQ (Moxifloxacin, Ofloxacin, Levofloxacin or fluoroquinolone in general). No: not MDR or XDR.

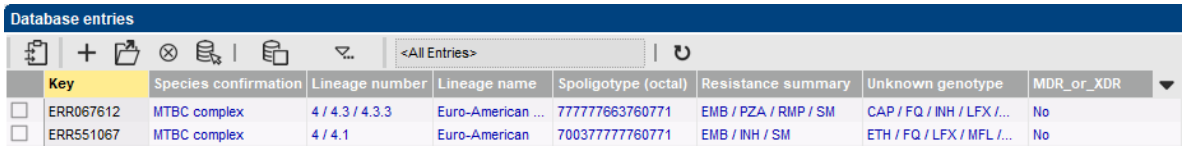

Figure 15: Example output of the Species, Lineage, Spoligotyping and Resistance information fields.

The character experiment types for **Spoligotype**, **Resistance** and **Genomic variants** are created and updated with the predicted traits. Please note that the shown names of the experiment types can be different in your case depending on whether you choose an alternative name during installation.

10. Open a character card for one of the analyzed entries by clicking on the corresponding green colored dot in the *Experiment presence* panel.

Below, the interpretation of the results gathered in the character experiment types is given.

Spoligo presence/absence experiment type: character experiment in which the binary code (for each of the 43 spacers: absence or presence) is stored (see Figure 16 and Figure 17).

Antibiotic resistance experiment type: character experiment containing all antibiotics from the selected resistance KB (see Figure 18).

The summary resistance call for an antibiotic can be (see Figure 19):

- ***Resistant (R; -5)***: at least 1 known resistance-related mutation or indel from the selected resistance KB was found with sufficient coverage.
- ***Unknown (U; -3)***: there are no known resistance-related mutations from the selected resistance KB found with sufficient coverage but there is at least 1 unknown mutation or indel found in a resistance-related gene. For positions already included in the resistance KB, every (non-synonymous) mutation or indel with sufficient coverage (as defined in the settings) leads to the 'unknown' status. For positions in resistance-related genes not yet included in the resistance KB, only majority/consensus mutations with more than 75% relative coverage lead to the 'unknown' status. Lastly, if no bases are covered in resistance related genes for a particular antibiotic, the outcome is also 'unknown'.
- ***Failed (F; -2)***: no known or unknown mutations with sufficient coverage were found but also not all positions in the resistance-related genes were covered.
- ***Susceptible (S; +1)***: there are no known or unknown mutations found and all bases of the resistance-related genes were covered sufficiently.

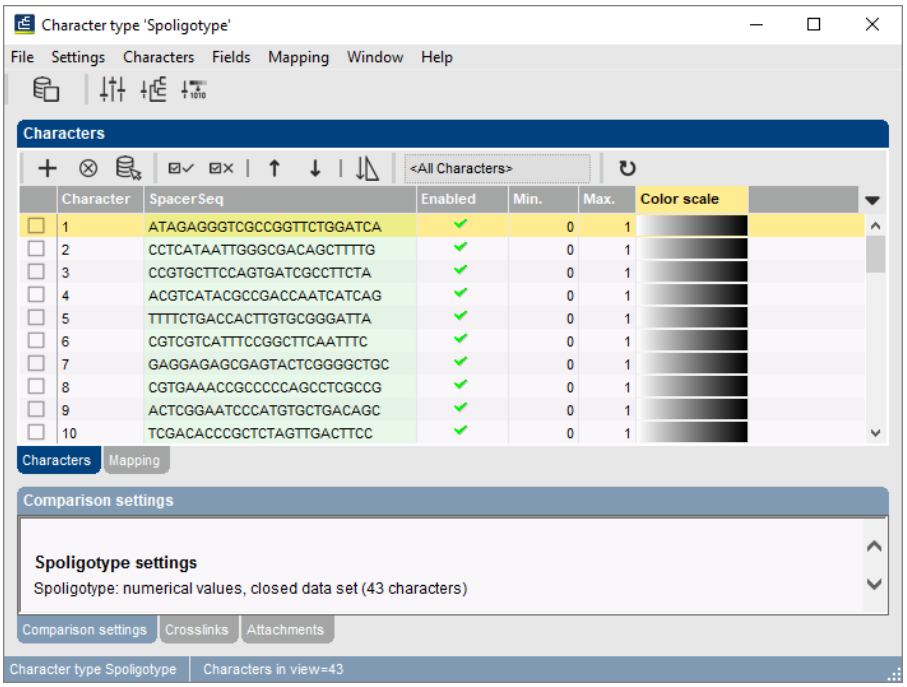

Figure 16: Spoligo presence/absence experiment type. This experiment contains 43 spacer sequences.

Figure 17: Example output of spoligotyping (spoligo presence/absence experiment type) for a sample with spoligotype 000000000003771. White = spacer is absent, black = spacer is present.

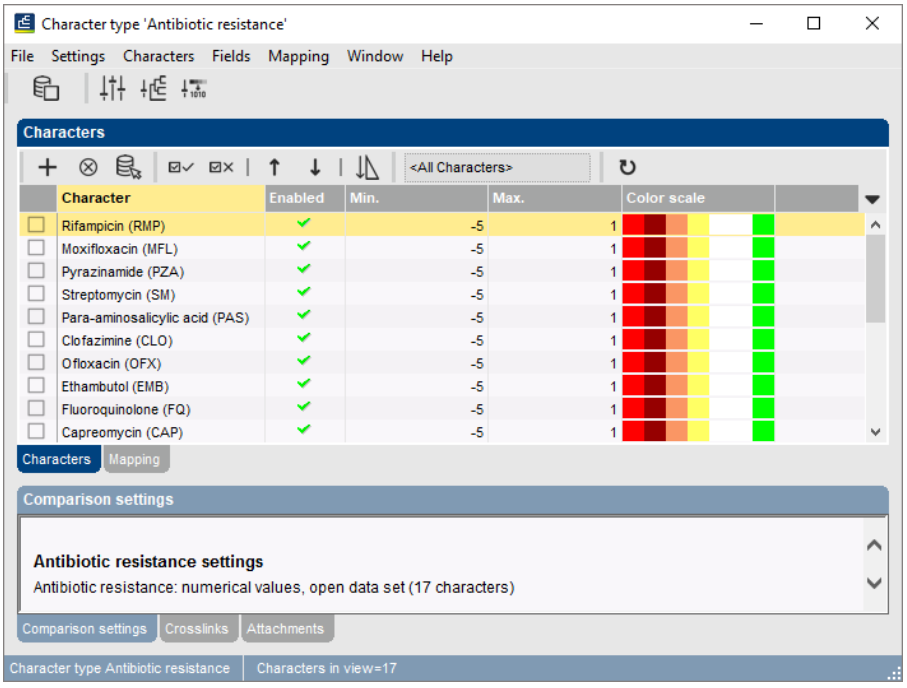

Figure 18: Antibiotic resistance experiment type.

Figure 19: Example output of antibiotic resistance experiment type for sample ERR067612. U = Unknown, R = Resistant, F = Failed, S = Susceptible.

Nucleotide variants experiment type: Character experiment that contains all nucleotide mutations in resistance-related genes that have sufficient coverage. These include (new) mutations which are not in the selected resistance KB (see Figure 20).

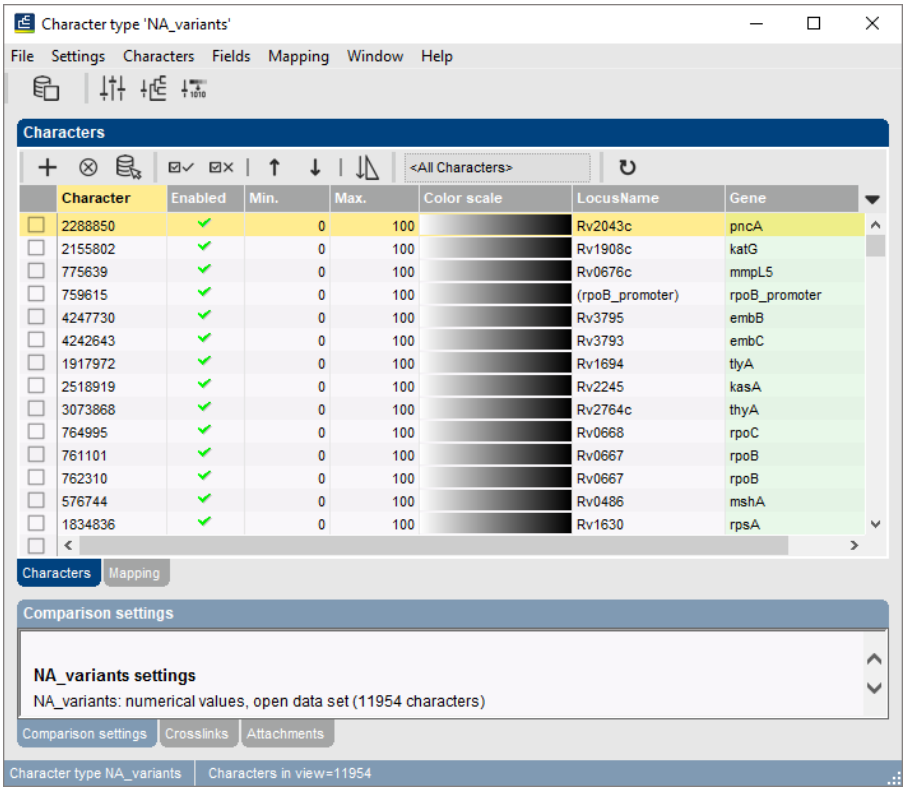

Figure 20: Nucleotide variants experiment type. Character: genomic position (H37Rv numbering).

The resistance call for a resistance-related position can be (Figure 21):

- A/C/T/G: nucleotide substitution.
- Multiple variants: there are multiple bases with sufficient coverage.

- Insertion: an insertion is present at this position.
- - : a deletion is present at this position.

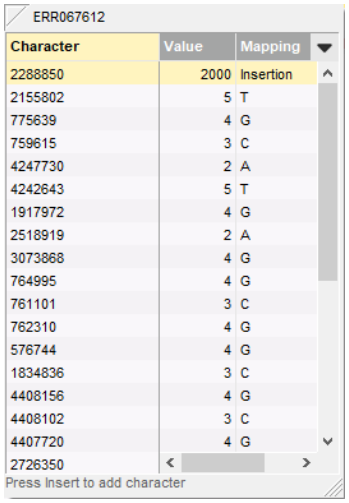

Figure 21: Example output of nucleotide variants experiment type for sample ERR067612.

Amino acid variants experiment type: Character experiment that contains all amino acid mutations in resistance-related genes that have sufficient coverage. These include (new) mutations which are not in the selected resistance KB.

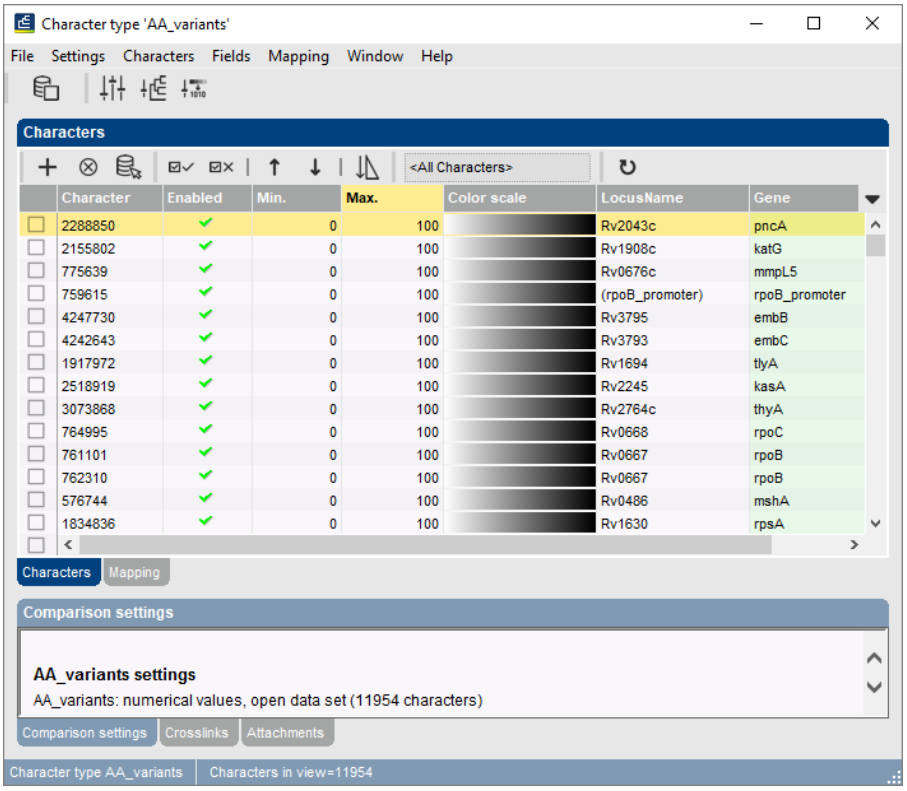

Figure 22: Aminoacid variants experiment type. Character: genomic position (H37Rv numbering).

The resistance call for a resistance-related position can be:

- An amino acid (one letter abbreviation).
- Multiple variants: there are multiple bases with sufficient coverage.
- Insertion: an insertion is present at this position.
- - : a deletion is present at this position.

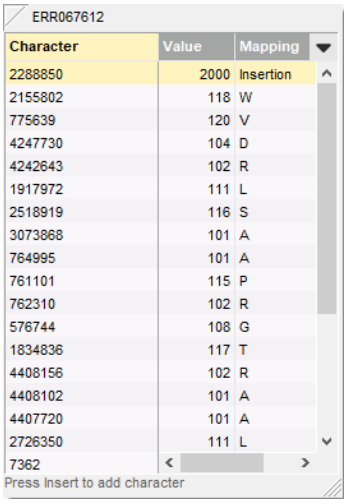

Figure 23: Example output of amino acid variants experiment type for sample ERR067612.

11. Close the character card(s) and click on the green colored dot corresponding to the Mapping experiment.

Mapping experiment type: Sequence experiment in which the mapped sequence of your sample will be stored. This sequence has the same length as the H37Rv reference genome (4411532 bp) (see Figure 24).

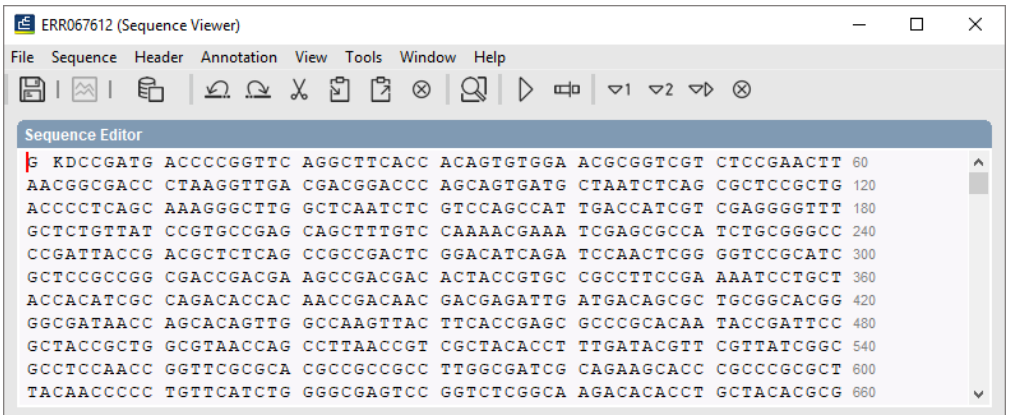

Figure 24: Example output of mapping experiment type for sample ERR067612.

12. Close the Mapping experiment type.

6 Reports

1. To generate a summary report, select the entries of interest and click on ***MTBC*** > ***Show report***.

The *Report* window contains a genotype report for each of the selected entries (see Figure 25).

2. Select another entry in the *Entries* panel to update the results in the *Genotype report* panel.

The creation date of the report (***Date***), the Key (***Name***), and information fields checked in the *Settings* are displayed in the *Genotype report* panel.

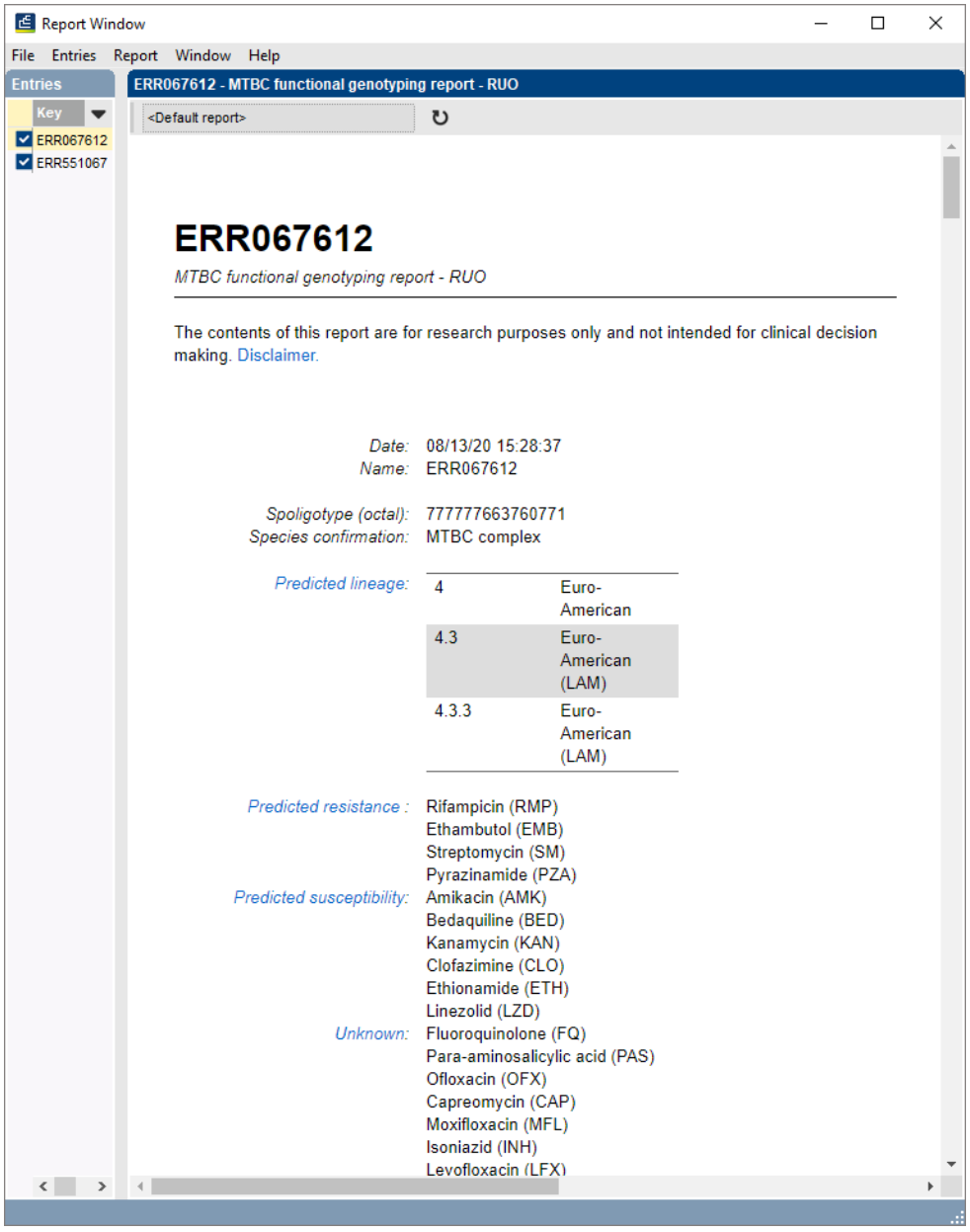

Figure 25: Example of a functional genotyping report.

3. To rerun an analysis from the *Report window* (with current MTBC settings) for one entry, click on ***Report*** > ***Reanalyze current report*** or click on ↻ at the top of the report.

4. To rerun the analysis for all entries displayed in the *Report window*, click on ***Entries*** > ***Reanalyze all***.

5. Click on a hyperlink of one of the predicted traits to display the detailed results in the *Genotype report* panel.

6. Select ***File*** > ***Exit*** to close the *Report* window.