BIONUMERICS Tutorial:

wgMLST typing: routine workflow starting from sequence read sets

1 Introduction

This tutorial explains how to prepare your database for wgMLST analysis and how to perform a full wgMLST analysis (de novo assembly, assembly-based and assembly-free calling) in BIONUMERICS on a routine basis starting from sequence read sets.

2 Installation of the plugin

1. Create a new database (see tutorial "Creating a new database") or open an existing database.
2. Call the *Plugins* dialog box from the *Main* window with ***File*** > ***Install / remove plugins...*** ( ).
3. Select the *WGS tools plugin* from the list in the *Applications tab* and press the <***Activate***> button.
4. Confirm the installation of the plugin.

The *Calculation engine URL* wizard page queries for the Uniform Resource Locator (URL) that uniquely identifies the calculation engine instance to connect to (see Figure 1).

With the ***Use default Cloud Calculation Engine*** option clients will use the Applied Maths cloud instance (`https://wgmlst.applied-maths.com`), which is hosted on Amazon servers in the US. This option should also be selected if you do not intend to run jobs on the calculation engine, but instead run all calculations on your own computer.

5. Make sure the ***Use default Cloud Calculation Engine*** option is selected and press <***Next***>.

In the *Organism and project* wizard page of the *WGS tools installation* wizard, two options are available (see Figure 2):

- Choose ***Local calculations only*** if you do not intend to run jobs on the calculation engine and instead wish to run all calculations on your own computer. With this option checked, an ***Organism*** should be selected from the drop-down list (do not select the ***No organism*** option if you wish to perform wgMLST). By selecting an organism, credentials to a demo project will be filled in automatically.
- Choose ***Enable running jobs on Cloud Calculation Engine*** to unlock the full potential of the default Cloud Calculation Engine. In this case, you will need credentials to your

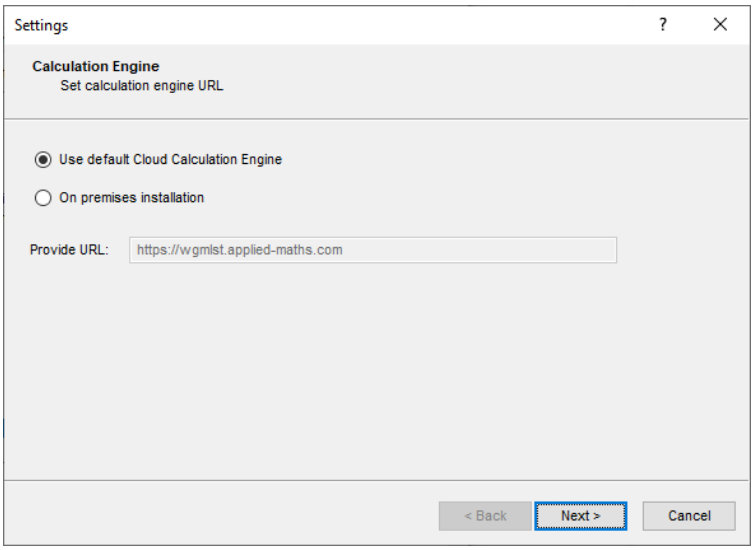

Figure 1: The *Calculation engine URL* wizard page in the *WGS tools installation* wizard.

own calculation engine project, for which credits can be purchased. Your ***Project name*** and corresponding ***Password*** should be entered in the corresponding text boxes. Pressing <***Request a new CE project***> will direct you to a form on the Applied Maths website where a new calculation engine project can be requested.

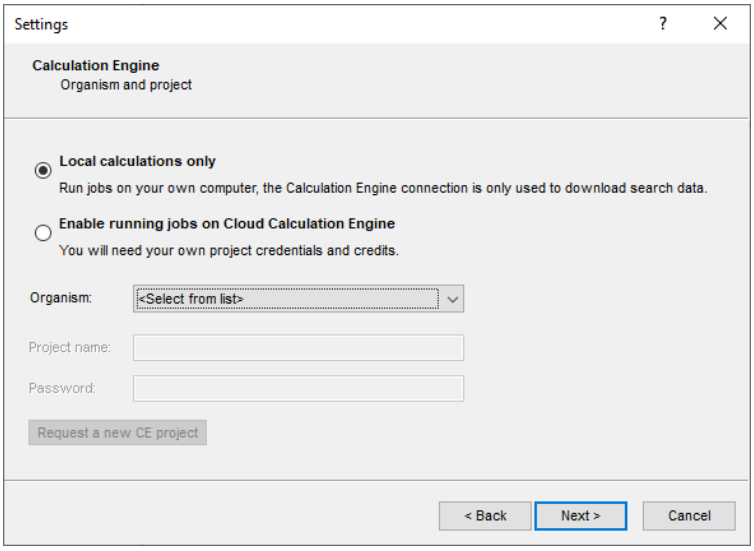

Figure 2: The *Organism and project* wizard page in the *WGS tools installation* wizard.

6. After having specified all correct settings, press <***Next***> to proceed with the installation.

BIONUMERICS will now download organism-specific settings and search data. A confirmation message pops up when the download is completed.

7. Press <***OK***> twice to finalize the installation of the plugin.

8. Press <***Exit***> to close the *Plugins* dialog box.

After installation of the plugin, the settings of the *WGS tools plugin* can be accessed with ***WGS tools*** > ***Settings...***.

9. Close and reopen the database to activate the features of the *WGS tools plugin*.

During installation of the plugin, the **wgMLST** character experiment is created and synchronized with the organism-specific locus scheme. All detected loci and subschemes are added to this experiment.

10. In the *Main* window double-click the character experiment type **wgMLST** in the *Experiment types* panel to call the *Character type* window.

11. Click on the drop-down bar in the toolbar (see Figure 3 for an example).

The views that have been defined at the curator level are synchronized upon installation and are listed. In most databases following views are defined by the curator: the default view **All loci**, the **Core loci** view, the **MLST** view for the traditional seven housekeeping loci, and the **wgMLST loci** view containing all loci except the ones present in the **MLST** view.

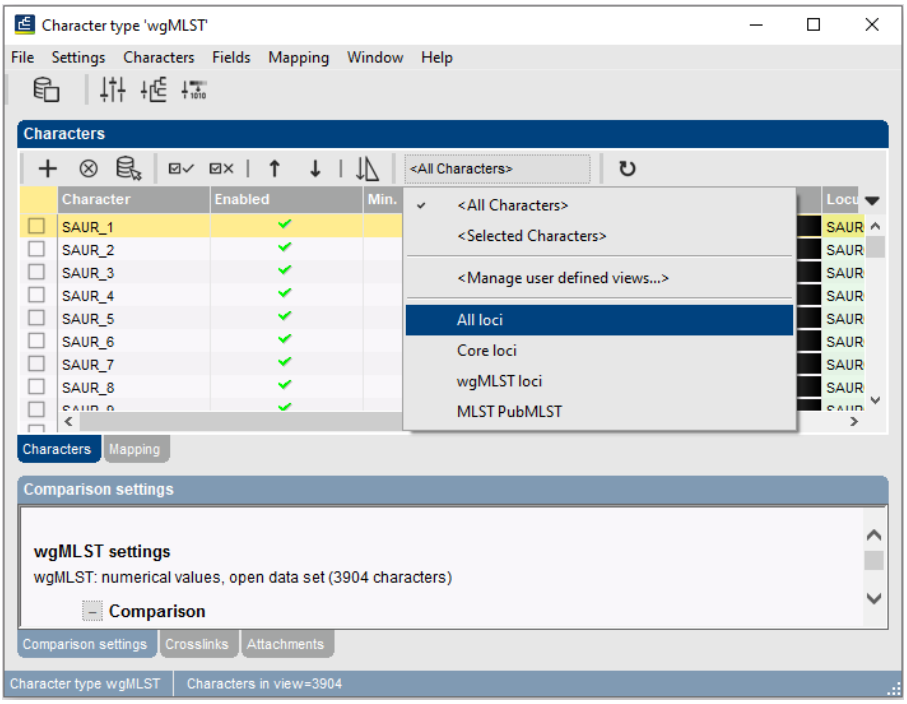

Figure 3: Views defined at the curator side.

12. Select another view from the list to update the set of loci in the *Characters* panel.

The number of loci in the selected view is displayed in the status bar at the bottom of the window.

13. To view all characters again, select <**All loci**> again from the drop-down list.

Besides these curator views, the user can create as many additional local character views as needed and use them as subscheme e.g. for clustering or when inspecting the allele calls for a subset of loci (select ***Characters*** > ***Character Views*** > ***Manage user defined views***).

14. Close the *Character type* window.

3 Import of sequence read sets

1. Select ***File*** > ***Import...*** (, **Ctrl+I**) to call the Import tree.

2. Click the +-sign next to the ***Sequence read sets data*** import option to display the sequence read sets import routines.

Two import routines are listed:

- ***Import sequence read sets***: With this option, a multitude of different file types can be imported and stored inside the database. We do not recommend to use this option since the files might fill up your BIONUMERICS database quickly and we want to avoid duplication of large data sets.
- ***Import sequence read set data as links***: With this option, only the link to the samples is stored in BIONUMERICS, resulting in a lightweight database. This option is only available after installation of the *WGS tools plugin* and is the preferred option when working with sequence read sets.

3. Make sure the ***Import sequence read set data as links*** option is selected in the Import tree and press <***Import***> (see Figure 4).

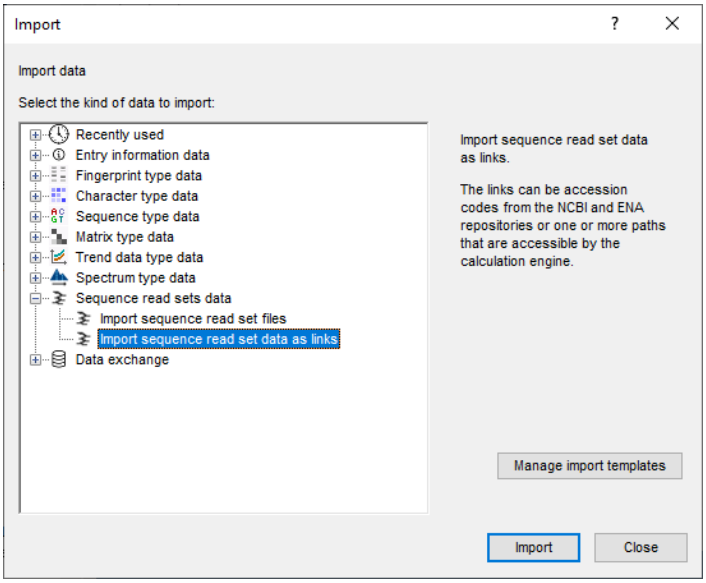

Figure 4: Import sequence read set data as links.

Links to multiple data sources are available, including online and offline data repositories such as: ***NCBI (SRA)***, ***EMBL-EBI (ENA)***, ***Amazon (S3)***, ***BaseSpace***, ***Alibaba OSS***, or ***Local file server*** (see Figure 5). Depending on the choice of import, different parameters may be queried in the next steps.

In this tutorial, the import from a local file server is covered. For more information about the other options, please consult the *WGS tools plugin* manual or the import tutorials on our website.

4. Select the ***Local file server*** and press <***Next***>.

5. Press the <***Browse***> button and select your `*.fastq` or `*.fastq.gz` files, located on your computer, external drive or on a network location (see Figure 6).

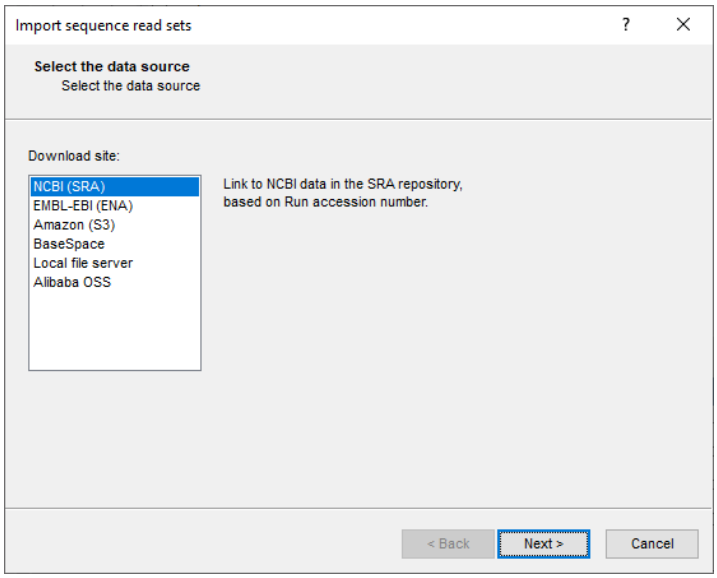

Figure 5: Data sources.

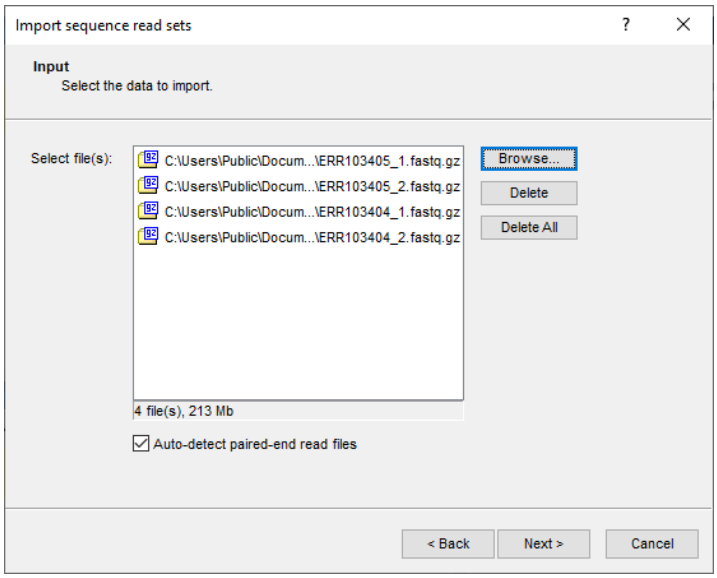

Figure 6: Select sample files.

The option ***Auto-detect paired-end files*** is default checked. This option ensures that the files are checked for the presence of paired-end data. Files that contain paired-end data are recognized by the same file name except for paired-end specific characters: e.g. same name apart from the _1 or _2 suffix. Below the file list, a brief summary on the selected files is displayed and updated. This summary indicates how many files of a specific file format were found, and their total file size.

6. Select <***Next***> to go to the next step.

A default import template is listed, parsing the sample names from the file names and linking the sample names to the ***Key*** field.

7. Press the <***Preview***> button to check the parsing based on the selected template.

8. Close the preview.

If the selected import template does not result in a correct parsing of your data, press the <***Edit***> button to change the rules (e.g. link the sample names to an entry information field) or press <***Create new***> to define a new template from scratch.

9. Make sure the import template and **wgs** experiment is selected and click <***Next***> to go to the next step (see Figure 7).

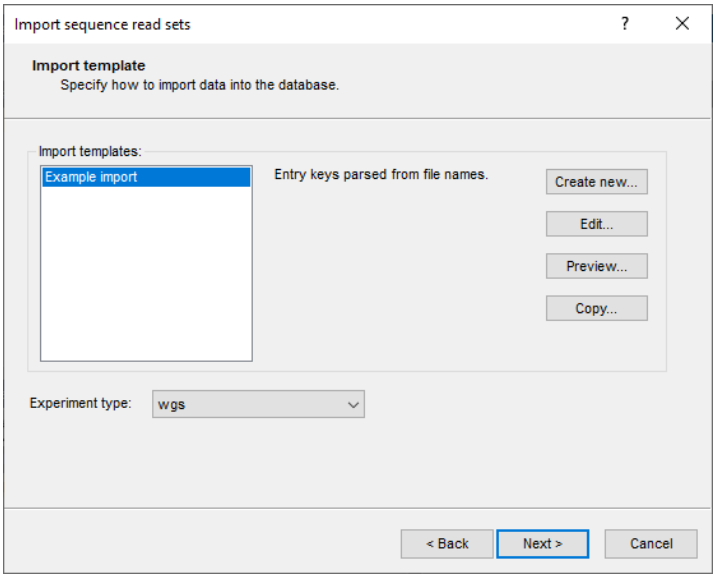

Figure 7: Import template.

The number of entries that will be created/updated during import is indicated.

10. Click <***Next***>.

In the last step, the wgMLST calculation jobs (de novo assembly, assembly-based and assembly-free calling) can be launched on the imported data links (***Open submit jobs dialog after import***). Note that same dialog can be called from the *Main* window at any time with ***WGS tools*** > ***Submit jobs...*** ( ) (see 4).

When the ***Local file server*** option was selected as data source, some basic statistics on the reads can be calculated upon import (***Calculate sequence read set statistics***). Based on the sequence read set statistics bad sequencing runs for which no jobs should be submitted to the calculation engine can be filtered out, saving you credits.

11. Make sure the ***Calculate sequence read set statistics*** option is selected, uncheck ***Open submit jobs dialog after import*** and press <***Finish***> to start the import of the data links.

Once the import is completed, the entries are created/updated and have one green dot next to it in the column of the sequence read set experiment type **wgs**.

12. Click on a green colored dot corresponding to the experiment type **wgs**.

The data links are displayed in the *Sequence read set experiment* window. If the option ***Calculate sequence read set statistics*** was checked in the last step, the statistics are displayed below (see Figure 9). Some statistics are also stored in the **quality** experiment and can be used to filter out bad sequencing results (see 4.1).

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

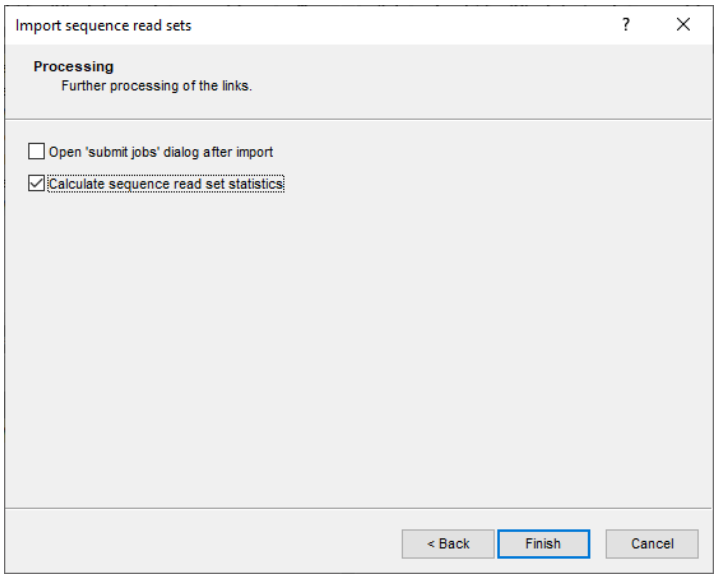

Figure 8: wgMLST calculation jobs.

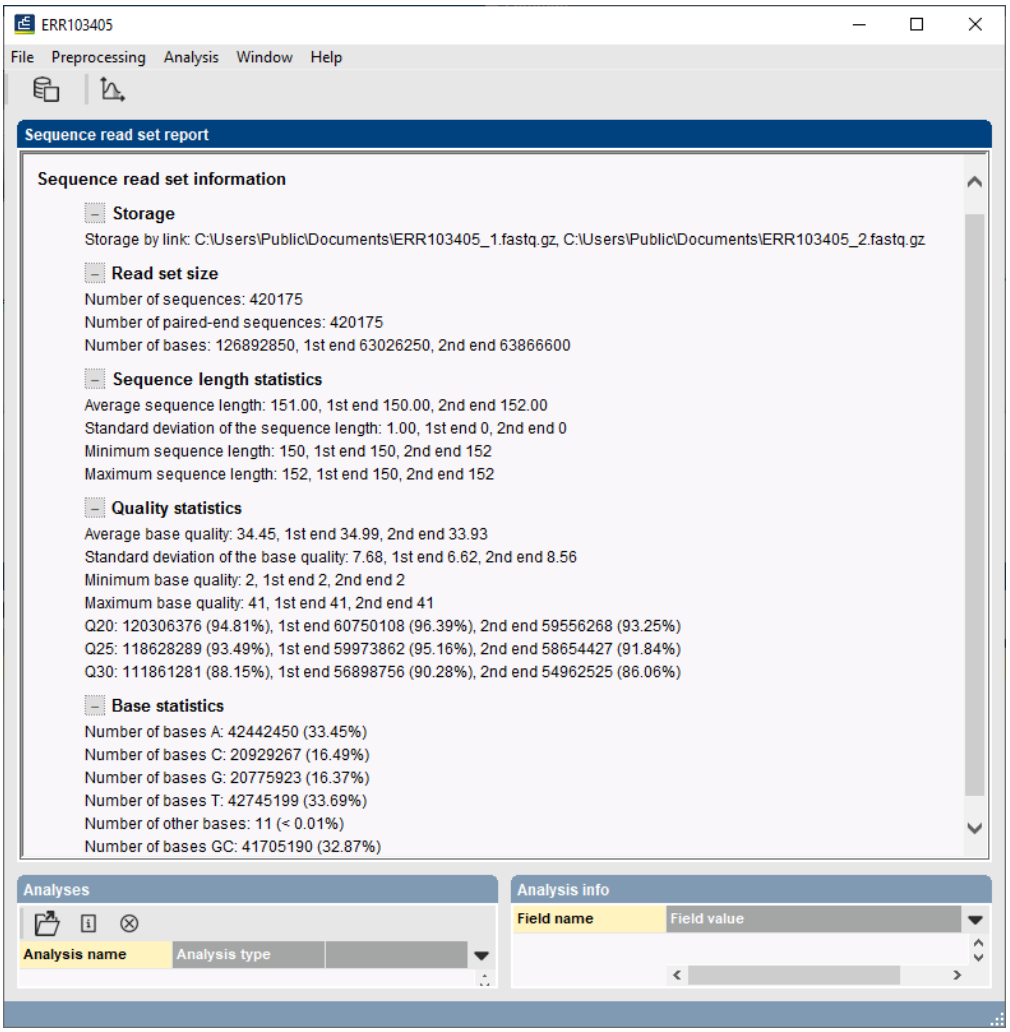

Figure 9: Sequence read set card.

4 Submission of jobs

4.1 Check read statistics

Before launching wgMLST jobs, it is recommended to take a look at the read statistics of the samples to filter out bad sequencing runs. These read statistics are calculated when the option ***Calculate sequence read set statistics*** was checked during import of the read set links (only available when importing reads from a ***Local file server***) and are saved in the **wgs** and **quality** experiments.

1. In the *Main* window, select the entries that you want to analyze using the check-boxes next to the entries or with the **Ctrl**- or **Shift**-keys.

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

3. Click on the next to the experiment name **quality** in the *Experiments* panel to display the quality data in the *Experiment data* panel.

4. Select ***Characters*** > ***Show values*** ( ) to show the corresponding character values for all entries in the comparison.

The quality values are displayed in the *Experiment data* panel (see Figure 10). The **AvgQuality** is an important indicator for the sequencing quality. The value depends on the sequencing technology used. For Illumina reads for example, the average read quality should be above 30. When samples have Illumina average quality values below 20, these should not be considered for further analysis.

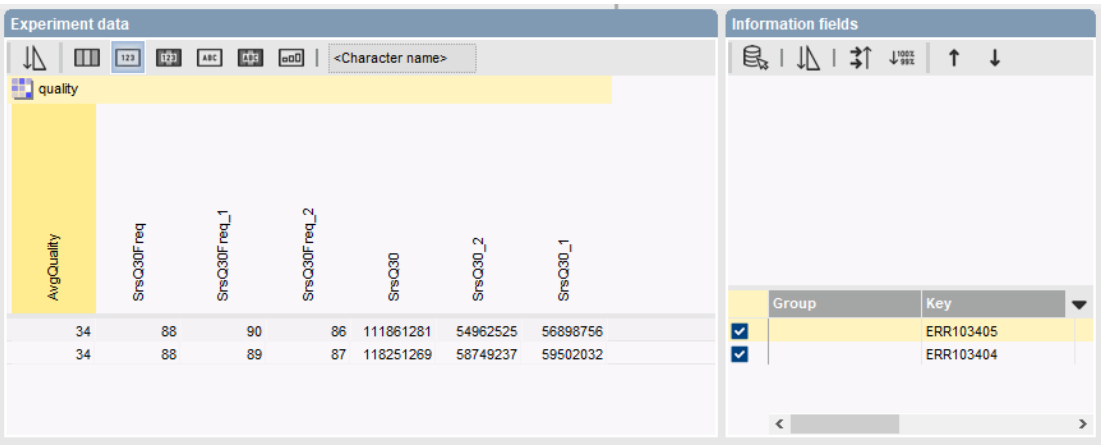

Figure 10: The quality experiment.

5. Close the *Comparison* window.

To make the distinction between good and bad sequencing run, the quality status (good or bad) can be entered in an entry information field:

6. To create a new entry field, make sure the *Database entries* panel is the active panel in the *Main* window, select ***Edit*** > ***Information fields*** > ***Add information field...***, specify a name (e.g. **Read quality**) and press <***OK***>. With the option ***Edit*** > ***Information fields*** > ***Edit field in selection...*** (**Ctrl+M**), text can be added to selected entries (e.g. “Good” or “Bad”).

7. Alternatively, you can opt to permanently remove entries with bad sequencing runs from the database: make sure the *Database entries* panel is the active panel in the *Main* window, select the entries you wish to remove and select ***Edit*** > ***Delete selected objects...*** (⊗).

4.2 Select jobs

Once the sequence reads sets are imported and linked to the ***wgs*** sequence experiment (see 3), the wgMLST jobs can be launched:

8. In the *Main* window, select the entries that you want to analyze using the check boxes next to the entries or with the **Ctrl**- or **Shift**-keys. Make sure that only samples with good quality reads are included in the selection (see 4.1 to check the good quality samples).

9. Select ***WGS tools*** > ***Submit jobs...*** ( ) to call the *Submit jobs* dialog box.

Alternatively check the ***Open submit jobs dialog after import*** option in the *Processing* wizard page during import of the data (see Figure 12).

In the *Submit jobs* dialog box you can define which algorithms can be run on the samples.

Following algorithms are available when starting from sequence read sets and when the ***Calculation Engine*** option is selected (see Figure 11):

- ***De novo assembly*** to calculate the de novo sequence assembly based on the reads retained after trimming. Following de novo assemblers are available on the cloud calculation engine: ***Velvet (Optimizer)***, ***SPAdes*** (default), ***SKESA*** and ***Unicycler***.
- ***wgMLST assembly-based calls*** to define the alleles based on a BLAST analysis on the de novo assembled contigs.
- ***wgMLST assembly-free calls*** to define the alleles directly from the reads.

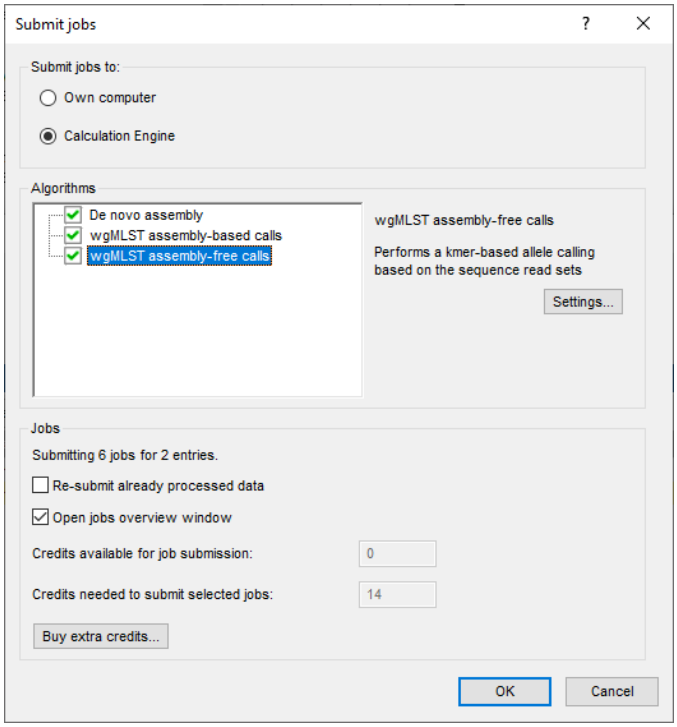

Figure 11: Calculation Engine - wgMLST jobs starting from sequence read sets.

Following algorithms are available when starting from sequence read sets and when the ***Own computer*** option is selected (see Figure 12):

- ***De novo assembly*** to calculate the de novo sequence assembly based on the reads retained after trimming. Only the ***SKESA*** de novo assembler is available.
- ***wgMLST assembly-based calls*** to define the alleles based on a BLAST analysis on the de novo assembled contigs.

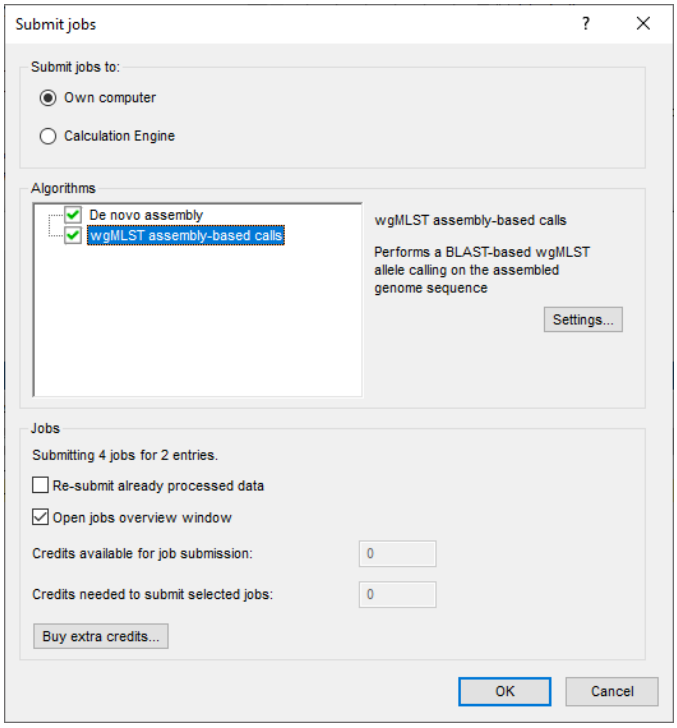

Figure 12: Local calculations - wgMLST jobs when starting from sequence read sets.

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

When the jobs are run on the ***Calculation Engine***, credits are required. Credit costs depend on the job that is submitted: 1 credit for de novo assembly, 3 credits for the assembly-based calls, and 3 credits for the assembly-free calls.

10. Check the algorithms that you wish to run on the samples, check (and optionally change) the settings, and press <***OK***> to launch the jobs.

When the jobs are run on the ***Calculation Engine*** and links are present to `*.fastq` or `*.fastq.gz` files stored on a local hard drive or a local file server a message will pop up asking to upload the files to an Amazon S3 temporary storage (called the **CE Store**), which the calculation engine can access (see Figure 13). Press <***OK***> to start the **CE Store Uploader** (see Figure 14).

4.3 Overview of the jobs

11. By default, the *Job overview* window will open after submission of the jobs. The same dialog can be called at any time with ***File*** > ***Jobs overview...*** ( ).

The *Submitted time*, the job *Status*, and much more can be monitored. In the *Message* field, the run comments are displayed in real time (see Figure 15).

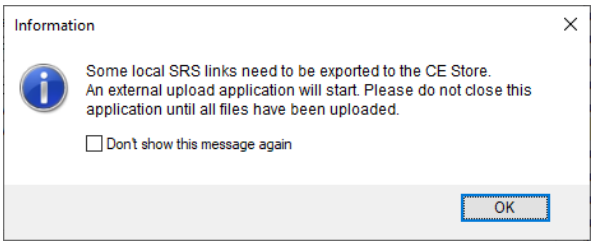

Figure 13: Upload to CE store.

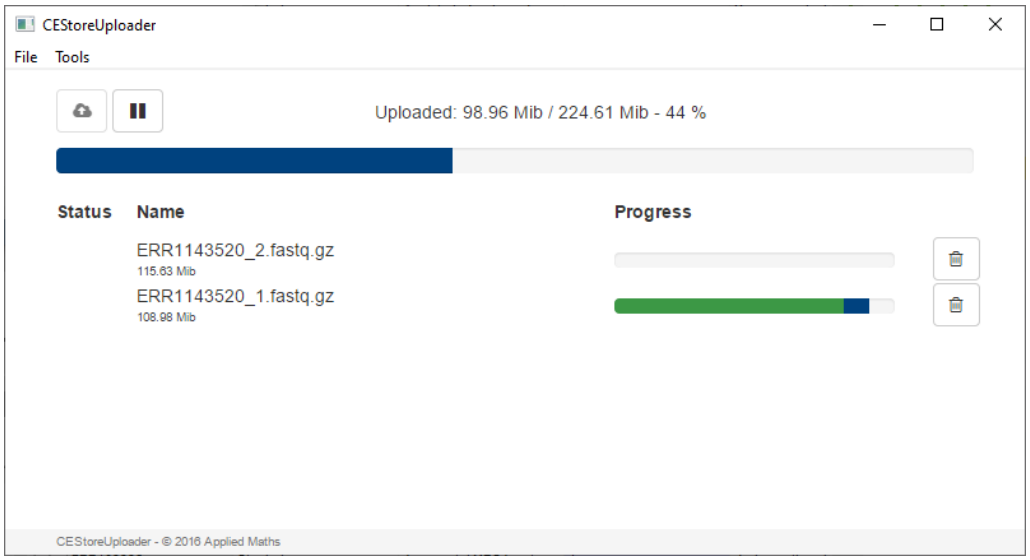

Figure 14: CE Store Uploader.

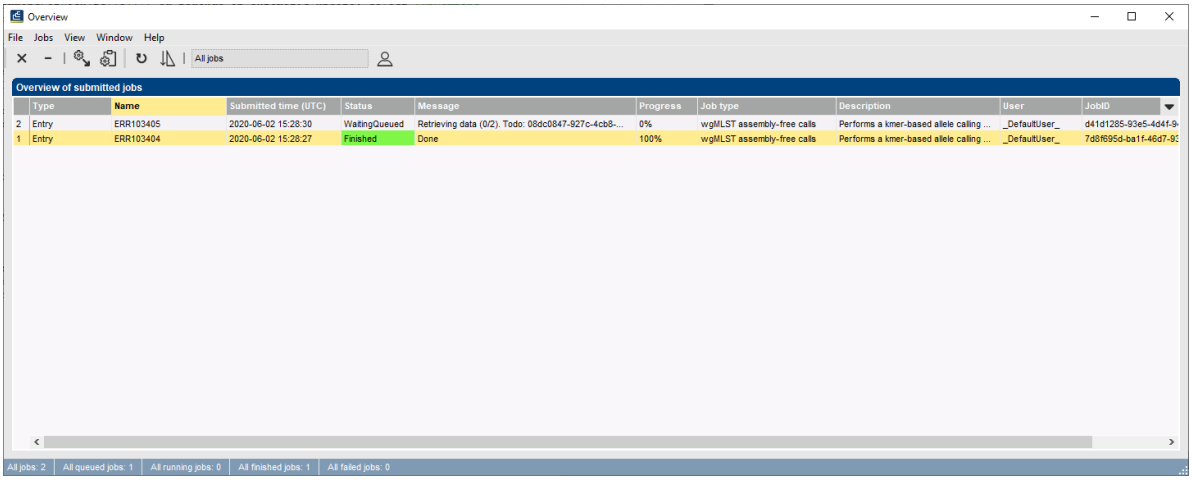

Figure 15: Job overview.

On average, the calculation time on the Calculation Engine for a novo assembly is around **20-30 min**; the assembly-free calling takes about **3 min**, and the assembly-based calling is finished after **5 minutes**. The calculation time of jobs running on your own computer depends on the computer hardware and is more difficult to estimate.

12. To refresh the overview, press ***View*** > ***Refresh*** (↻ , **F5**).

5 Job results

5.1 Import job results

There are two options available in the *Job overview* window to import the job results in your BIONUMERICS database:

1. 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:

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

All available job results (for the selected entries) will be imported to the database and linked to their respective entry and experiment type.

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.

Depending on the jobs that were checked in the *Submit jobs* dialog box, following information is stored in the BIONUMERICS database:

- When the option ***De novo assembly*** was checked, the results from the de novo assembly algorithm, i.e. concatenated de novo contig sequences with coverage information are stored in the sequence experiment type **denovo**.
- The **wgMLST** experiment contains the allele calls for the detected loci, where the consensus from assembly-based and assembly-free calling - if both jobs were submitted - resulted in a single allele ID. In case multiple allele calls are made and different calls obtained for the same locus, default the lowest common allele ID is retained for these loci (select ***WGS tools*** > ***Settings...*** to access this setting in the *wgMLST tab*).
- The character experiment type **quality** contains the quality statistics for the raw data and algorithms that were applied.
- The sequence read set experiment type **wgs_TrimmedStats** contains some data statistics about the reads that were retained after trimming and that were used for the de novo assembly and the assembly-free calling.

5.2 Check job results

The character experiment type **quality** provides insight in the quality of the reads (see 4.1) and the results obtained for the different submitted jobs. The possible presence of low quality reads, assemblies and contaminations can be consulted in a very quick and easy way in the *Comparison* window.

3. In the *Database entries* panel of the *Main* window, select the entries that you want to analyze using the check-boxes next to the entries or with the **Ctrl**- or **Shift**-keys.

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

5. Click on the next to the experiment name **quality** in the *Experiments* panel to display the quality data in the *Experiment data* panel.

6. Select ***Characters*** > ***Show values*** ( ) to show the corresponding character values for all entries in the comparison.

7. Click on the drop-down list next to the **quality** experiment in the *Experiments* panel to display the default defined character views (see Figure 16).

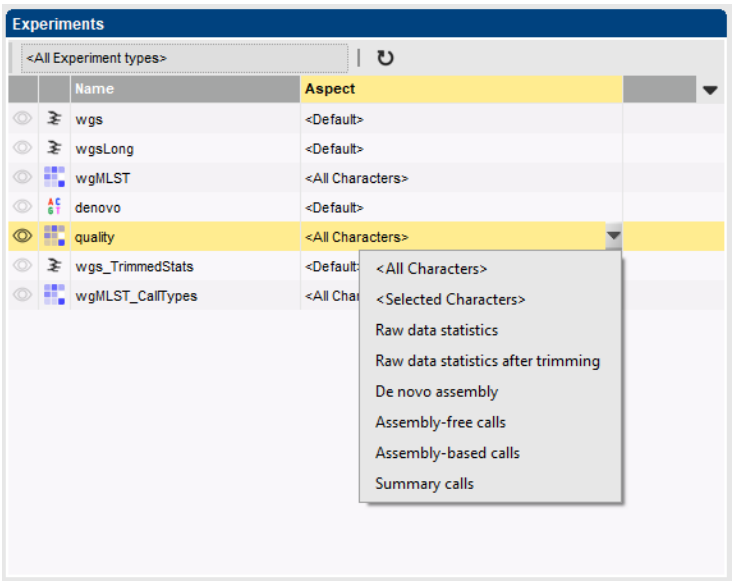

Figure 16: Character views.

The quality parameters are grouped based on the data sets and algorithms and the view can be restricted to each of these groups: raw data statistics (after trimming) (***AvgQuality***, ***Srs***, etc.), de novo assembly (***N50***, ***NrContigs***, etc.), assembly-free calls (***NrAF***), assembly-based calls (***NrBAF***), and summary calls (***NrConsensus***).

Some parameters are more informative and important than others. A few initial parameters for a first check are listed below:

- **AvgQuality**: the average quality depends on the sequencing technology used. For Illumina reads, the average read quality should be above 30.
- **AvgReadCoverage**: the expected coverage for each base is calculated based on the number of bases in the reads and the expected sequence length. Samples with coverages below 15 should be removed from the analysis. Ideally this number should be above 30.
- **Length**: this length should be close to the length you expect for your organism. Assemblies that are a lot smaller than expected, can be removed from the analysis. For larger lengths, it depends on the cause (contamination or presence of a plasmid).
- **NrAFMultiple**: some loci might have multiple allele hits so a low number is acceptable. If a very high number of multiple allele hits is observed, this indicates the presence of contamination.
- **CorePercent**: this parameter is only present when a core scheme has been defined for the organism. The acceptable range depends on the organism, with typical ranges between 95 and 100. For more diverse organisms, a lower percentage is acceptable, for clonal organisms it is not. This parameter also depends on how strict the core was defined and the

diversity of the strains used to define the core. Only very low numbers should be removed without further investigation.

Optionally, you can restrict the view in the *Experiment data* panel to only those parameters that are of interest to you (see Figure 17 for a custom character quality view):

8. Select the columns in the *Experiment data* panel that you want to include in your custom view while holding the **Ctrl-** key, double-click the **quality** character experiment in the *Experiment types* panel in the *Main* window, select ***Characters*** > ***Character Views*** > ***Manage user defined views...*** (<All Characters>), press the <***Add***> button, specify a name and select the ***Subset based*** option.

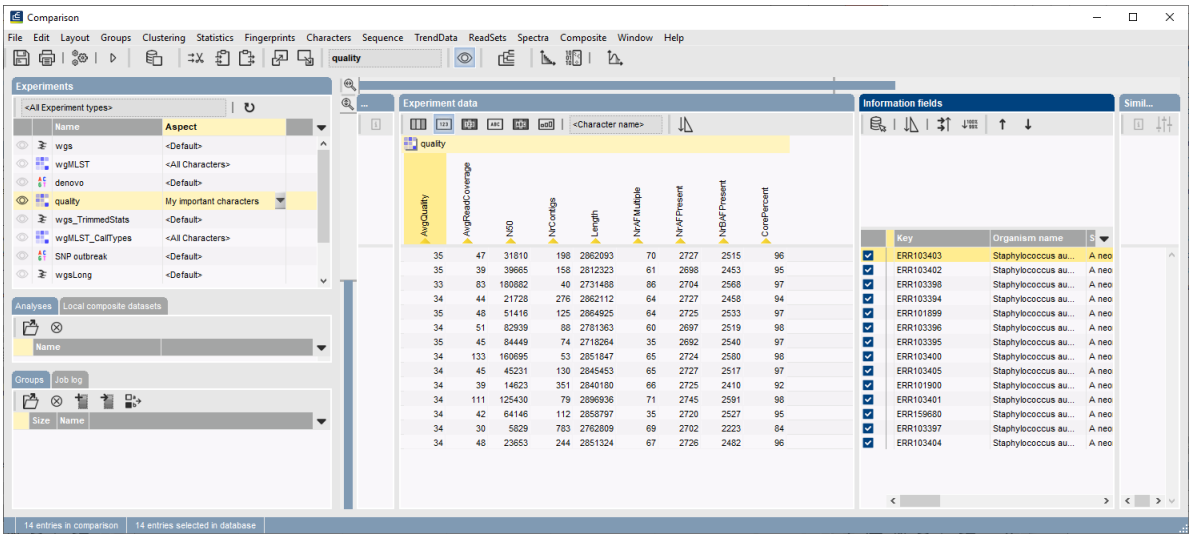

Figure 17: The *Comparison* window displaying a selection of quality parameters.

9. Close the *Comparison* window.

To make the distinction between good and bad job results, this status can be entered in an entry information field:

10. To create a new entry field, make sure the *Database entries* panel is the active panel in the *Main* window, select ***Edit*** > ***Information fields*** > ***Add information field...***, specify a name (e.g. **Job results**) and press <***OK***>. With the option ***Edit*** > ***Information fields*** > ***Edit field in selection...*** (**Ctrl+M**), text can be added to selected entries (e.g. “Good” or “Bad”).

11. Alternatively, you can opt to permanently remove entries with bad sequencing runs from the database: make sure the *Database entries* panel is the active panel in the *Main* window, select the entries you wish to remove and select ***Edit*** > ***Delete selected objects...*** (⊗).

6 Follow-up analysis

6.1 Comparison window

A cluster analysis on the **wgMLST** character experiment (or a subscheme thereof) is created in the *Comparison* window or the *Advanced cluster analysis* window.

1. In the *Database entries* panel of the *Main* window, select the entries that you want to analyze using the check-boxes next to the entries or with the **Ctrl**- or **Shift**-keys.

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

3. Click the drop-down list in the **Aspect** column of the **wgMLST** character experiment in the *Experiments* panel.

All subschemes defined by the curator in the allele database and the schemes defined by the user (if any) are listed (see Figure 18 for an example). One can very easily switch between the different aspects.

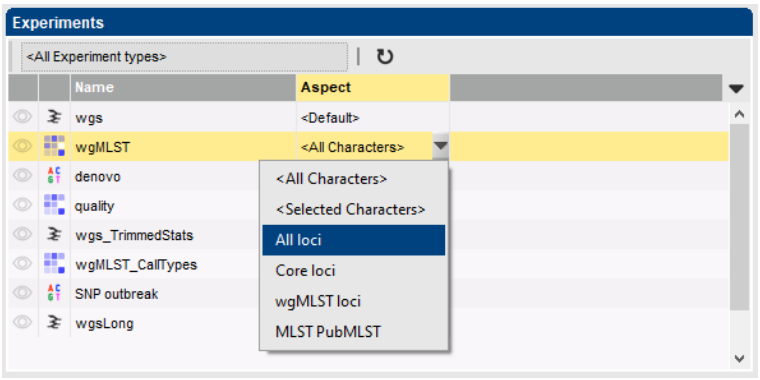

Figure 18: Character views.

A few analysis tools are highlighted in this tutorial that can be applied on wgMLST data:

- Similarity based clustering (see 6.2).
- Minimum spanning tree (see 6.3).

6.2 Similarity based clustering

4. Make sure the correct subscheme of the **wgMLST** character experiment that you want to use for your analysis (e.g. ***wgMLST loci***, ***Core loci***) is selected in the *Experiments* panel.

5. In the *Experiments* panel click on the eye icon ( ) that proceeds **wgMLST** to display the values of the selected aspect.

6. In case of closely related isolates select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...*** and choose the ***Categorical (differences)*** coefficient from the list (see Figure 19).

The ***Categorical (differences)*** coefficient treats each different value as a different state, and results in a distance matrix. With the ***Scaling factor*** one can deal with the hard-coded maximum of 200 that can be calculated for a distance value. Values that make sense are 1, 10 and 100, allowing the correct visualization of maximally 200, 2000 and 20000 different character values, respectively, in a cluster analysis.

7. Press <***Next***>, choose ***Complete Linkage*** in the last step and press <***Finish***>.

When the maximum distance of 200 has been reached, a message is displayed (see Figure 20). To avoid clipping of the dendrogram, repeat the previous steps and increase the ***Scaling factor*** with 10 or 100.

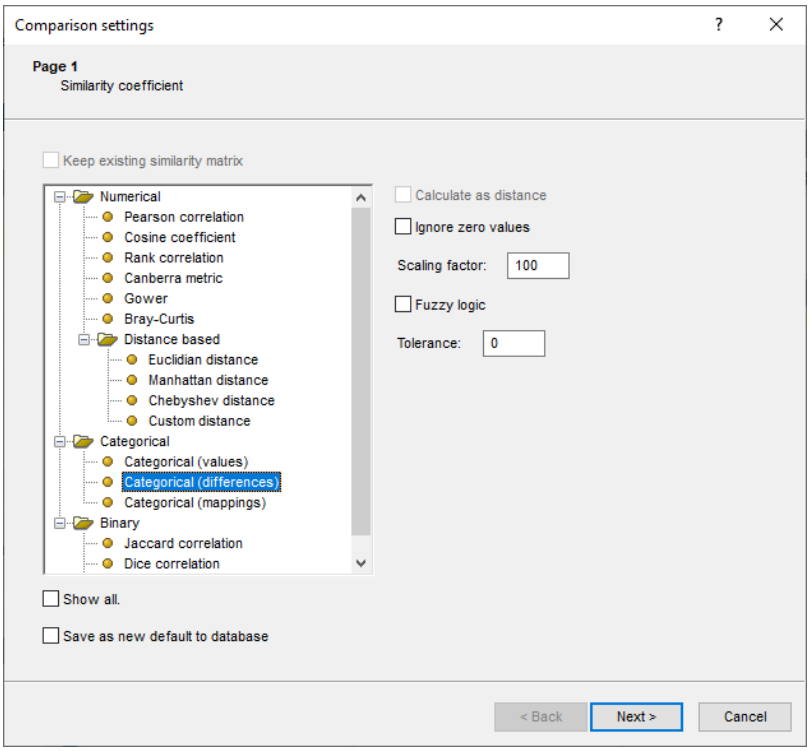

Figure 19: Similarity coefficients.

Figure 20: Maximum number.

The resulting dendrogram is displayed in the *Dendrogram* panel and the analysis is stored in the *Analyses* panel. The subscheme that was used is indicated between brackets: e.g. **wgMLST(Core loci)**.

8. The settings used to calculate the dendrogram that is displayed in the *Dendrogram* panel can be called with ***Clustering*** > ***Show information*** ( ).

9. To view the number of allele differences on the branches, select ***Clustering*** > ***Dendrogram display settings...*** ( ), and tick the option ***Show node information*** (see Figure 21).

To trace back the number of different loci from the branches or distance matrix, the displayed values needs to be multiplied with the ***Scaling factor*** used.

10. The polymorphic loci for the set of samples in the selected scheme can be displayed with ***Characters*** > ***Filter characters*** > ***Select polymorphic characters...***.

11. The information displayed in the *Experiment data* panel can be exported with ***Characters*** > ***Export character table***. The character table will open as a `export.csv` file in MS Excel.

12. To export the cluster analysis as it appears in the *Comparison* window select ***File*** > ***Print preview...*** (, **Ctrl+P**). The *Comparison print preview* window appears.

More features present in the *Comparison* window are explained in the BIONUMERICS manual.

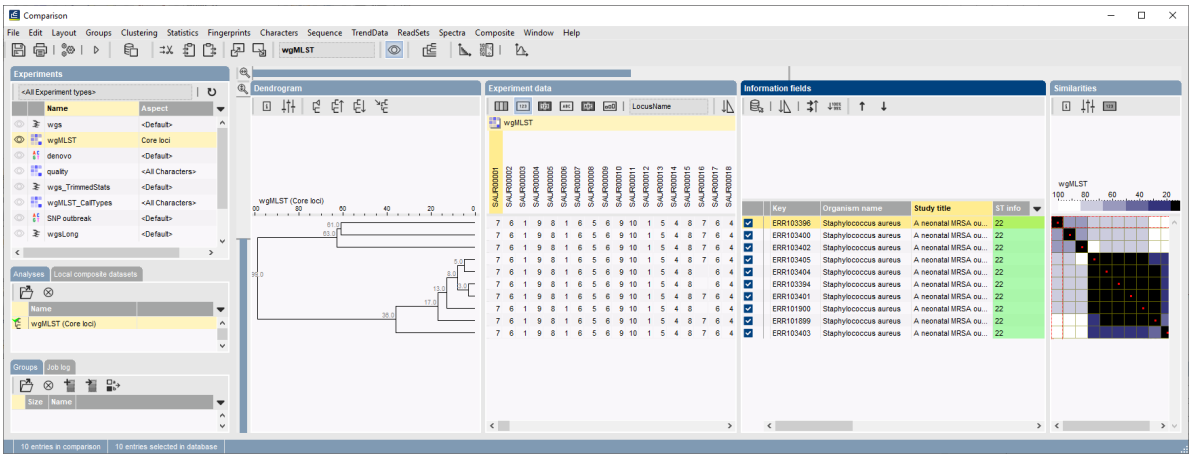

Figure 21: Complete linkage tree based on categorical differences.

6.3 Minimum spanning tree

A minimum spanning tree is calculated in the *Advanced cluster analysis* window which is launched from the *Comparison* window.

13. Select ***Clustering*** > ***Calculate*** > ***Advanced cluster analysis...*** in the *Comparison* window 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.

14. Specify an analysis name, make sure the correct subscheme is selected, select ***MST for categorical data***, and press <***Next***>.

To view and modify the settings of a selected template check the option ***Modify template settings for new analysis***.

A MST is now computed in the *Advanced cluster analysis* window (see Figure 22). The *Network panel* 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. The analysis is also added to the *Analyses* panel in the *Comparison* window.

15. Press or choose ***Display*** > ***Display settings*** to open the *Display settings* dialog box.

16. In the *Branch labels and sizes panel*, you can specify that you want to see the distances between the nodes (i.e. the number of allele differences): check ***Show branch labels*** and set ***Number of digits*** to “0”.

17. Click <***OK***> to close the *Display settings* dialog box. The MST is now displayed with branch labels.

More features present in the *Advanced cluster analysis* window are explained in the BIONUMERICS manual.

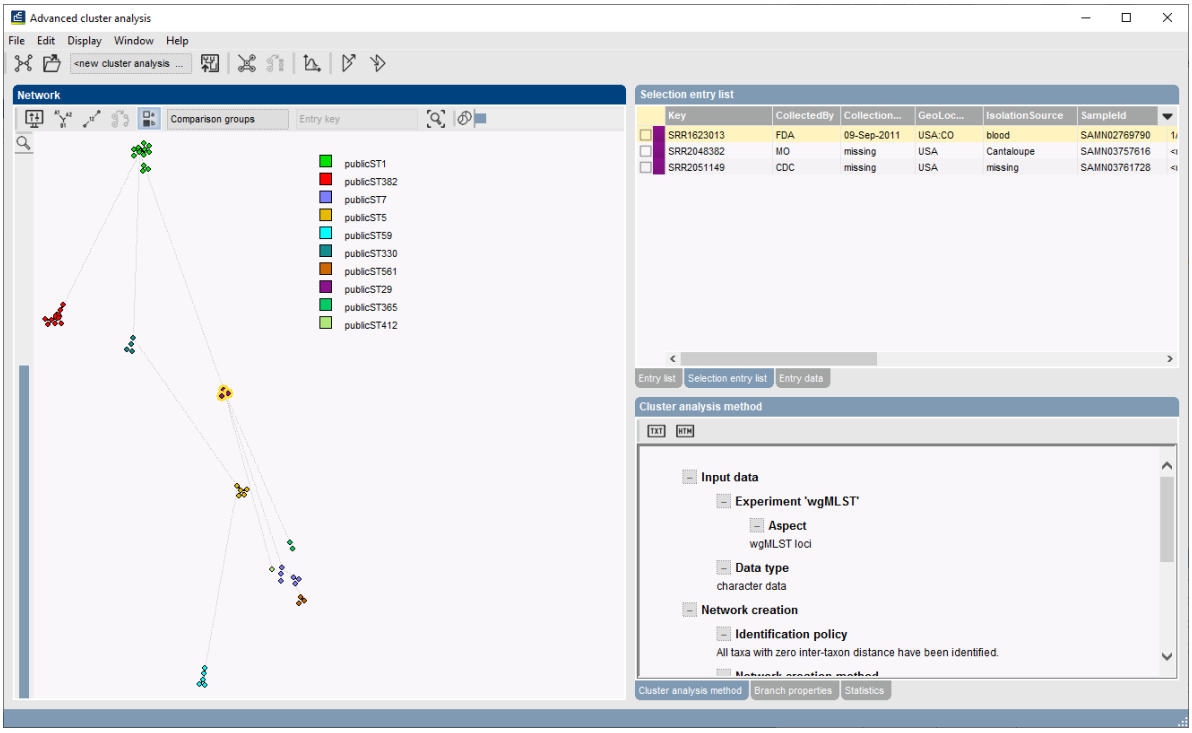

Figure 22: The *Advanced cluster analysis* window.