BioNumerics Tutorial:

wgMLST typing in BioNumerics: routine workflow starting from imported genomes

1 Introduction

This tutorial explains how to prepare your database for wgMLST analysis and how to perform a wgMLST analysis starting from assembled genomes that are imported in BioNumerics.

The installation of the *WGS tools plugin* requires credentials obtained from Applied Maths. Please make sure you have the credentials ready when following the steps in this tutorial. After installation of the plugin you will notice that importing and analyzing assembled genomes is a very easy and intuitive process.

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.

In the first step, the settings for the connection to the **calculation engine** need to be defined. The demanding calculations, here only the allele calling starting from the assembled genomes, will be performed on this external calculation engine. The user has the option to use the calculation engine present at the Applied Maths Amazon cloud instance (**Applied Maths cloud**) or to connect to a locally installed instance (**On premises**) (see Figure 1).

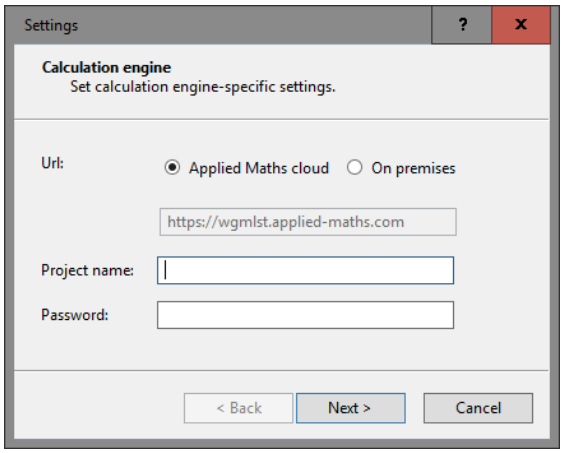

Figure 1: The *Calculation engine* wizard page in the *WGS tools installation* wizard.

5. Select the correct calculation engine resource. In most cases this will be the **Applied Maths cloud** option.
6. Specify the project name as obtained from Applied Maths. The project name is linked with the available credits for a specific account.

7. Specify the password that is used in conjunction with the specified project.

8. Press <***Next***> to proceed to the second step.

The organism (group) can be picked from the drop-down list with available organism schemes. The number of loci is indicated (see Figure 2 for an example).

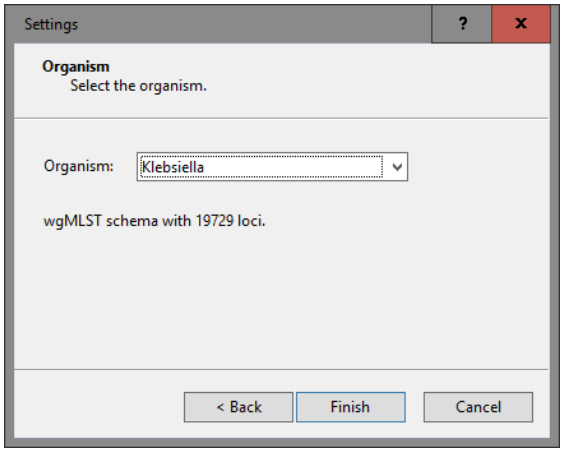

Figure 2: The *Organism* wizard page in the *WGS tools installation* wizard.

9. Press <***Finish***> to start the synchronization with the specified allele database.

The synchronization process can take a couple of minutes, depending on the number of loci and subschemes present in the allele database. A confirmation dialog is displayed when the synchronization has been completed (see Figure 3 for an example).

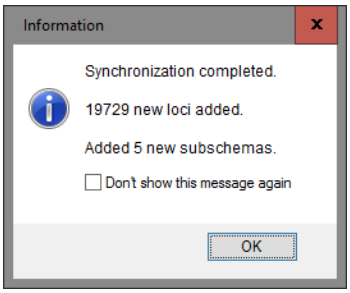

Figure 3: Confirmation of successful installation.

10. Press <***OK***> and <***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...***.

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.

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

12. Click on the drop-down bar in the toolbar (see Figure 4 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,

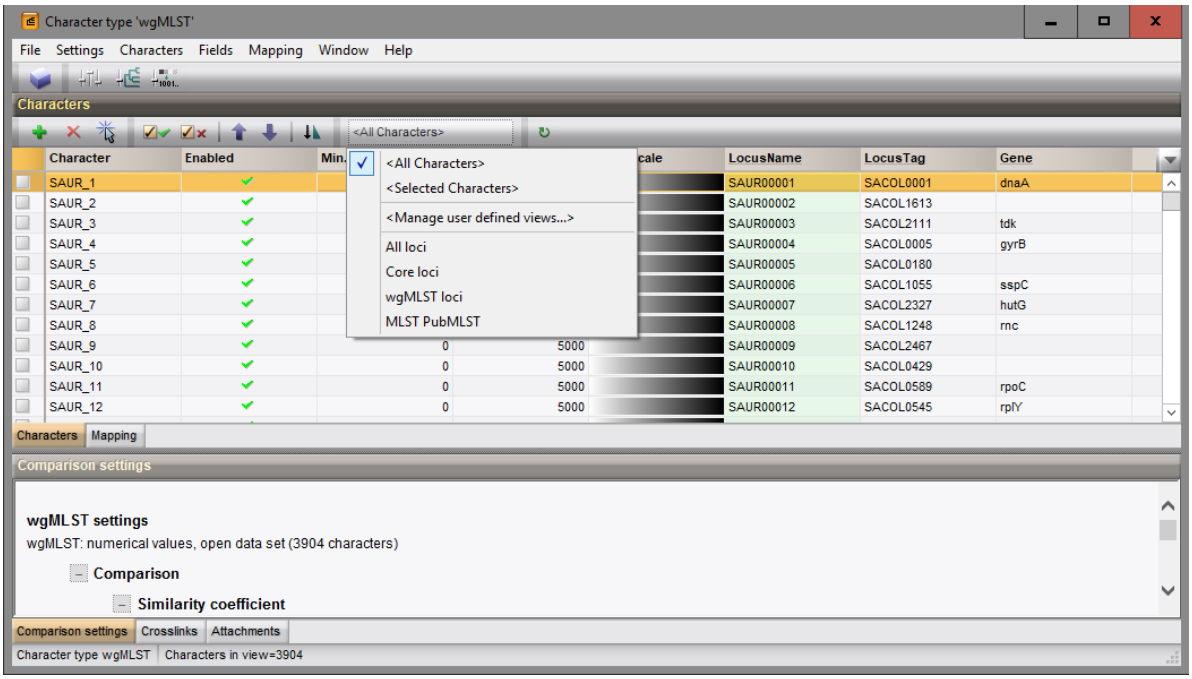

Figure 4: Views defined at the curator side.

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.

13. 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.

14. 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***).

15. Close the *Character type* window.

3 Import of assembled genomes

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

The sequence import routines are grouped in the Import tree under ***Sequence data***.

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

Following routines are available for the import of assembled genome sequences (see Figure 5):

- ***Import FASTA sequences from text files***: Sequences in FASTA format can be imported from text formatted files and linked to new or existing database entries. Optionally, FASTA tags can be stored in entry information fields.
- ***Import EMBL/GenBank sequences from text files***: Sequences in EMBL or GenBank format can be imported from text formatted files and linked to new or existing database entries. Header and feature

descriptions are automatically stored with the sequences. Optionally, EMBL and GenBank tags can be stored in entry information fields.

- ***Import sequence assemblies from BAM files***: A sequence assembly can be imported in BAM or SAM format using this import routine.
- ***Download sequences from internet***: The EBI, NCBI and NIG/DDBJ databases can be accessed directly from the BioNumerics software to import genome sequences in the BioNumerics database.

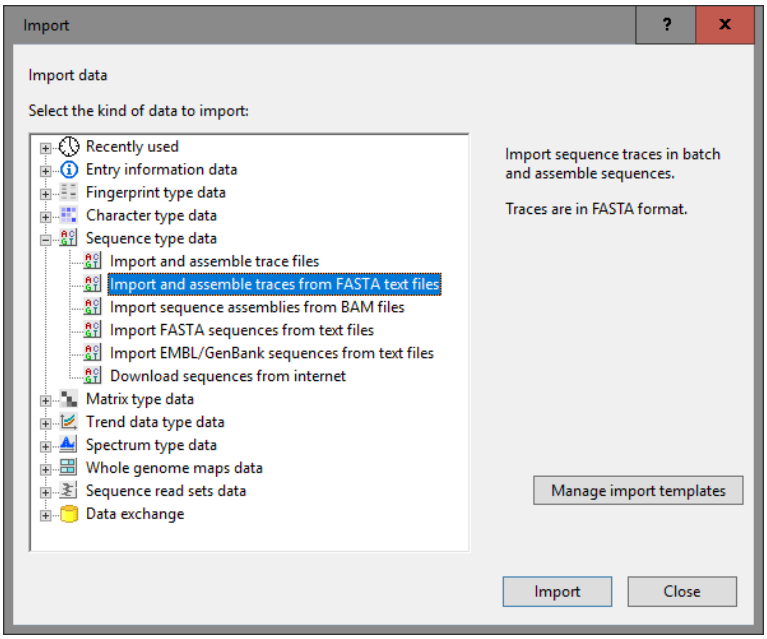

Figure 5: The Import tree.

In this tutorial, the import routines will not be covered. For detailed information about each of these import routines, please check the BioNumerics reference manual or the corresponding tutorials and movies on our website.

Note that when importing sequences using one of these import routines, the assembled genomes should preferably be linked to the ***de novo*** sequence experiment in one of the last steps of the Import wizard, since this experiment is defined as the default sequence experiment upon installation of the *WGS tools plugin*.

4 Submission of jobs

4.1 Select jobs

After import of the genome sequences in the ***de novo*** sequence experiment, the assembly-based calling job can be launched:

1. In the *Main* window, select the entries that you want to analyze using the checkboxes next to the entries or with the **Ctrl**- or **Shift**-keys.
2. Select ***WGS tools*** > ***Submit jobs...*** ( ) to call the *Submit jobs* dialog box.

In the *Submit jobs* dialog box you can define which algorithms can be run on the samples (see Figure 6). When only genome sequences are linked to the selected entries (and no read sets), only the ***Assembly-based***

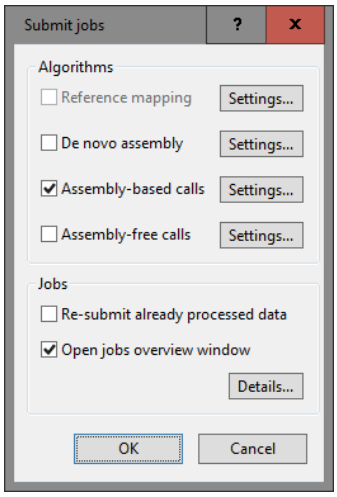

Figure 6: Assembly-based calls.

calls algorithm will be available. This algoritm will define the alleles based on a BLAST analysis on the (de novo assembled) genomes.

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.

Credit costs depend on the job that is submitted: 3 credits are required for the assembly-based job. For more information on the credits press the <***Details***> button.

3. Check the ***Assembly-based calls*** algorithm, check (and optionally change) the settings, and press <***OK***> to launch the jobs on the calculation engine.

4.2 Overview of the jobs

4. By default, the *Calculation engine overview* window will open after submission of the jobs. The same dialog can be called at any time with ***WGS tools*** > ***Jobs overview...*** ( ).

The *Entry* key, the *Submitted time*, the job *Status*, a *Description* of the job and its *Progress* and much more can be monitored. In the *Message* field, the run comments are displayed in real time (see Figure 7).

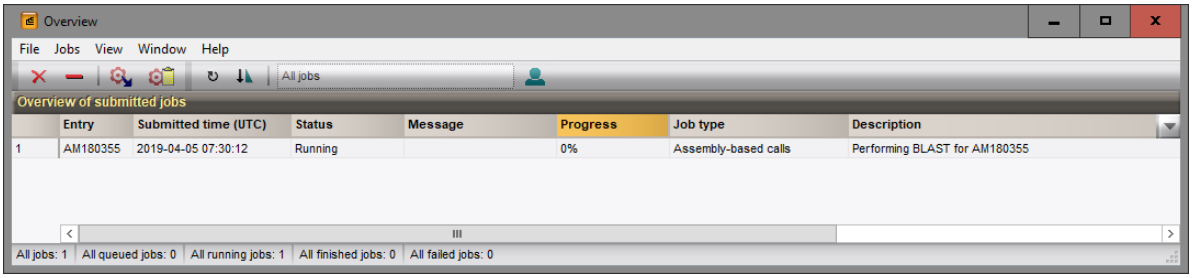

Figure 7: Job overview.

On average, the calculation time for an assembly-based calling is **8 to 9 minutes**.

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

5 Job results

5.1 Import job results

There are two options available in the *Calculation engine 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.

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 *Calculation engine overview* window.

For the ***Assembly-based*** jobs following information is stored in the BioNumerics database:

- The **wgMLST** experiment contains the allele calls for the detected loci.
- The character experiment type **quality** contains some quality statistics (see 5.2).

5.2 Check job results

The character experiment type **quality** provides insight in the results obtained for the submitted jobs.

3. 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.

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 8).

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), de novo assembly, assembly-free calls, assembly-based calls , and summary calls. When only an assembly-based job was launched, only the **De novo assembly**, **Assembly-based calls** and **Summary calls** views will contain data:

De novo assembly

- ***N50 (N50)***: Length of the median contig (in terms of sequence length).
- ***Contigs (NrContigs)***: The number of contigs in the sequence.

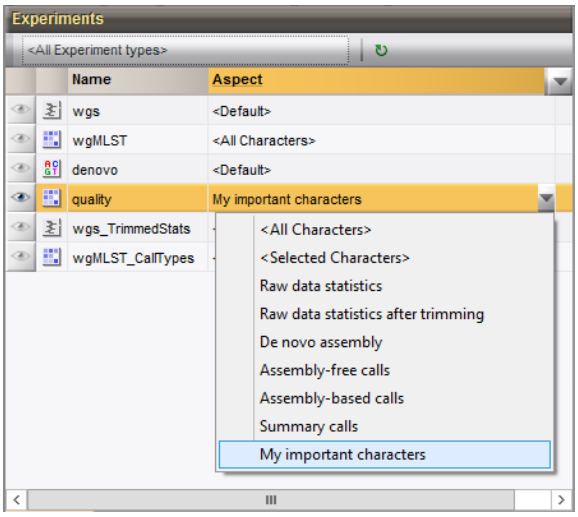

Figure 8: Character views.

- ***Sequence length (Length)***: Length of the sequence.
- ***Bases ACGT (NrBasesACGT)***: Number of bases A, C, G, and T.
- ***Bases N (NrBasesN)***: Number of bases N.
- ***Bases non ACGTN (NrNonACGT)***: Number of ambiguous bases (not taking N bases into account).

Assembly-based calls

- ***Multiple alleles (NrBAFMultiple)***: Number of loci with multiple allele hits. The preferred allele hit is the one with the lowest allele number.
- ***Perfect matches (NrBAFPerfect)***: Number of loci with at least one known allele hit that is 100% identical to an approved allele in the curator database.
- ***Alleles to submit (NrToBeSubmitted)***: Number of loci with an allele hit eligible for submission to the curator database. Only allele hits with a sequence identity of at least a user-specified threshold (and less than 100%) and whose sequence contains only non-ambiguous bases can be submitted.
- ***Submitted alleles (NrAlreadySubmitted)***: Number of loci which have already been submitted to the curator database.
- ***Present alleles (NrBAFPresent)***: Number of loci with at least one allele hit (= perfect (100%) matches and non-perfect matches). Must be close to the expected number of loci for the organism as defined by the curator.

8. Close the *Comparison* window.

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 *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 9 for an example). One can very easily switch between the different aspects.

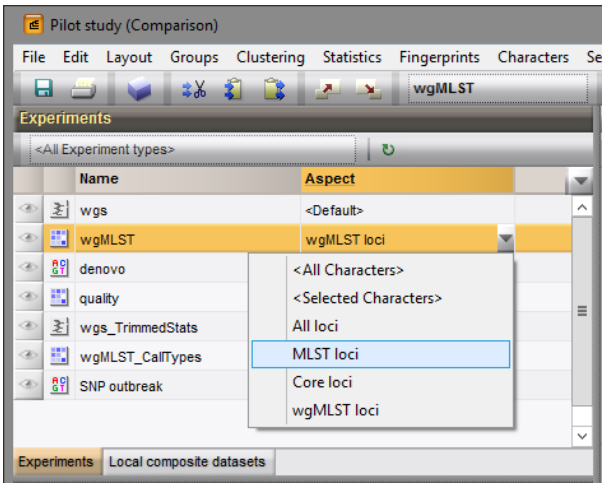

Figure 9: Character views.

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

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 10).

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***>.

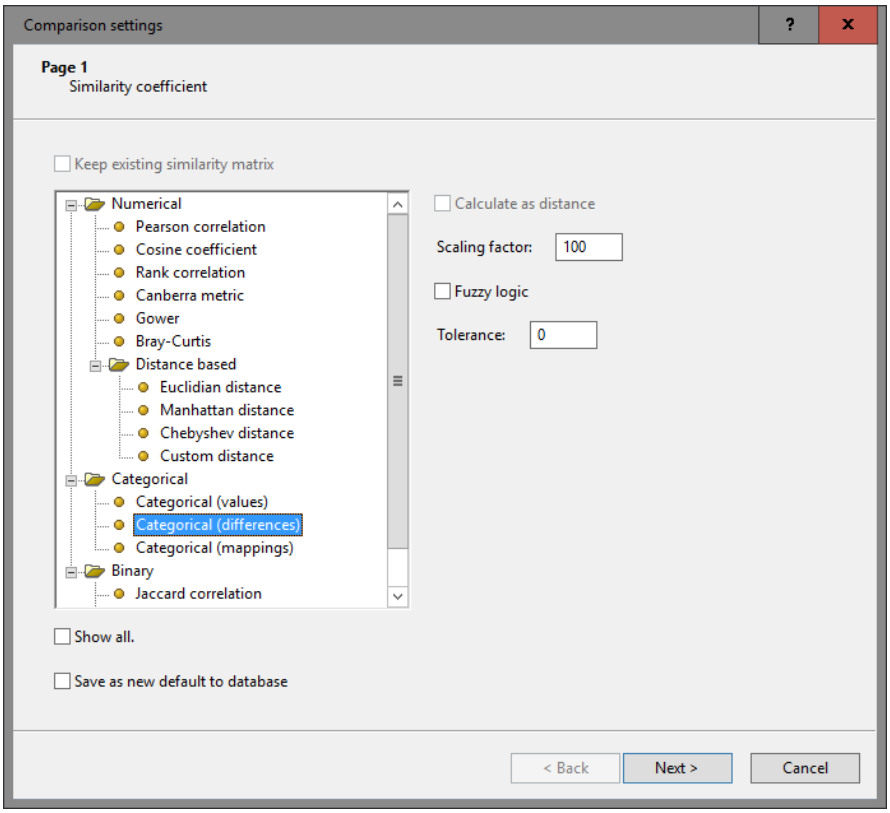

Figure 10: Similarity coefficients.

When the maximum distance of 200 has been reached, a message is displayed (see Figure 11). To avoid clipping of the dendrogram, repeat the previous steps and increase the ***Scaling factor*** with 10 or 100.

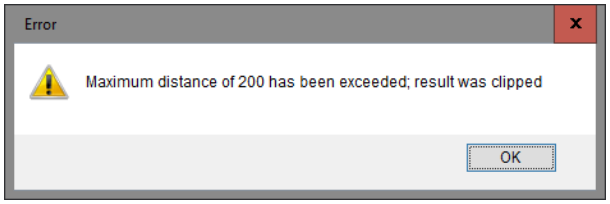

Figure 11: Maximum number.

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 12).

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.

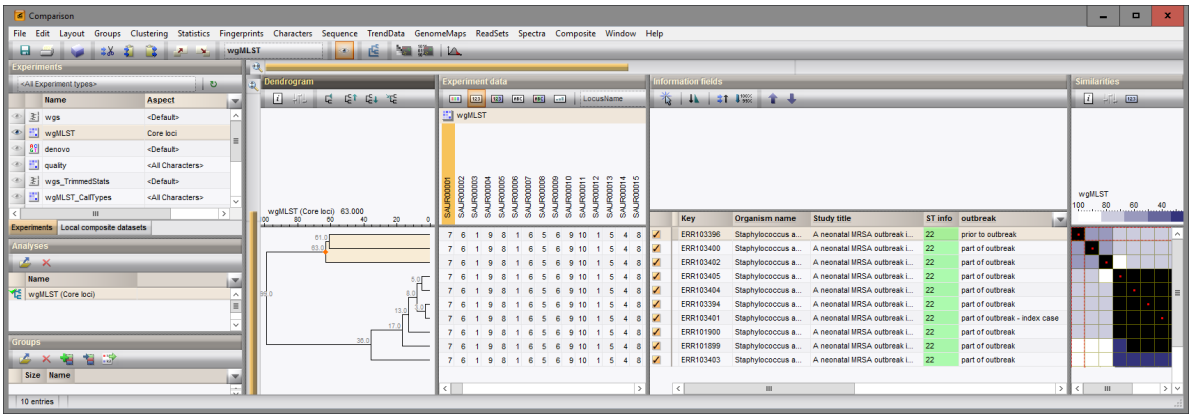

Figure 12: Complete linkage tree based on categorical differences.

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.

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 13). 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.

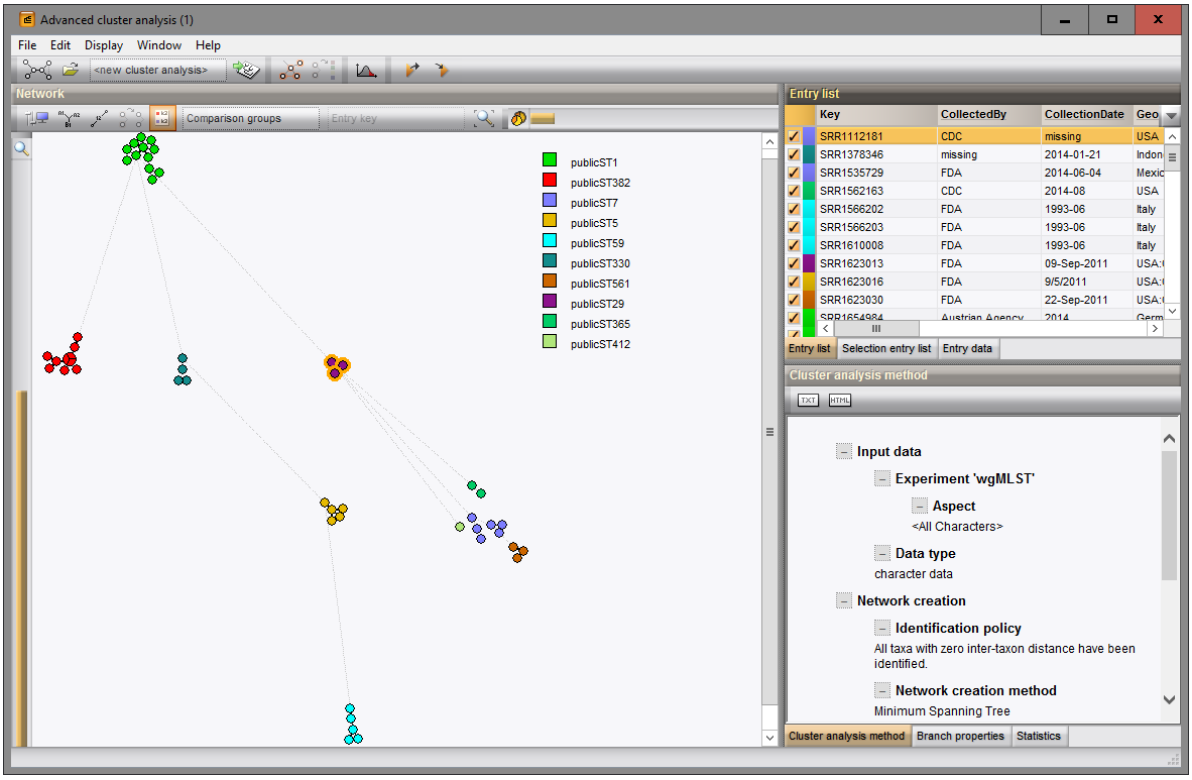

Figure 13: The *Advanced cluster analysis* window.