BIONUMERICS Tutorial:

wgMLST typing in the *Brucella* demonstration database

1 Introduction

This guide is designed for users to explore the wgMLST functionality present in BIONUMERICS without having to post calculation jobs on their own computer or on the external calculation engine. The whole genome demonstration database used in this tutorial contains the results obtained from the full wgMLST analysis in BIONUMERICS on publicly available sequence read sets and genome sequences of *Brucella* spp.

Although this guide provides the necessary information to start working with the wgMLST functionality present in BIONUMERICS, it is recommended to read the following documentation available for download on the tutorial page on our website:

- Tutorial "wgMLST typing: routine workflow starting from sequence read sets"
- Tutorial "wgMLST typing: routine workflow starting from imported genomes"
- Tutorial "wgMLST typing: detailed exploration of results"
- *WGS tools plugin* manual

Furthermore, a leaflet on our website (https://www.applied-maths.com/sites/default/files/extra/Brucella-spp-how-to-make-the-most.pdf) explains how to use the trunk and branch wgMLST schema for *Brucella* spp. efficiently.

2 Preparing the demonstration database

The **WGS demo database** for *Brucella* can be downloaded directly from the *BIONUMERICS Startup* window (see 2.1), or restored from the back-up file available on our website (see 2.2).

2.1 Option 1: Download demo database from the Startup Screen

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

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

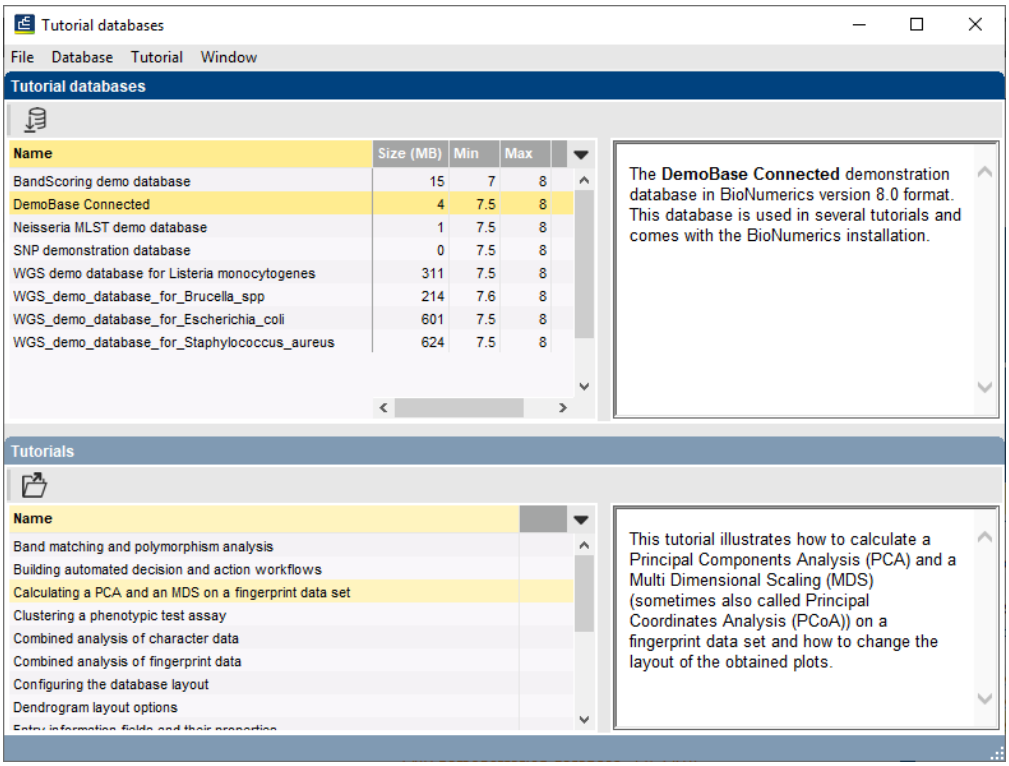

Figure 1: The *Tutorial databases* window, used to download the demonstration database.

2. Select the **WGS_demo_database_for_Brucella_spp** from the list and select ***Database*** > ***Download*** ( ).

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

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

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

5. Double-click the **WGS_demo_database_for_Brucella_spp** in the *BIONUMERICS Startup* window to open the database.

2.2 Option 2: Restore demo database from back-up file

A BIONUMERICS back-up file of the WGS demo database for *Brucella* is also available on our website. This backup can be restored to a functional database in BIONUMERICS.

6. Download the file `WGS_BRU.bnbk` file from https://www.applied-maths.com/download/sample-data, under 'WGS_demo_database_for_Brucella_spp'.

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

8. From the menu that appears, select ***Restore database...***.

9. Browse for the downloaded file and select ***Create copy***. Note that, if ***Overwrite*** is selected, an existing database will be overwritten.

10. Specify a new name for this demonstration database, e.g. “WGS Brucella demobase” (see Figure 2).

11. Click <***OK***> to start restoring the database from the backup file.

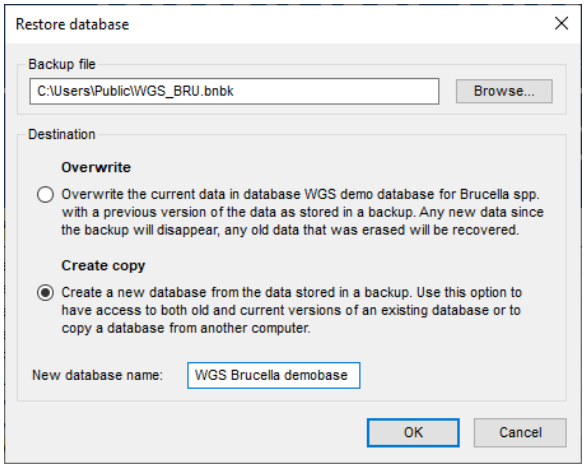

Figure 2: Restore *Brucella* demonstration database from backup file.

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

The *Main* window is displayed (see Figure 3).

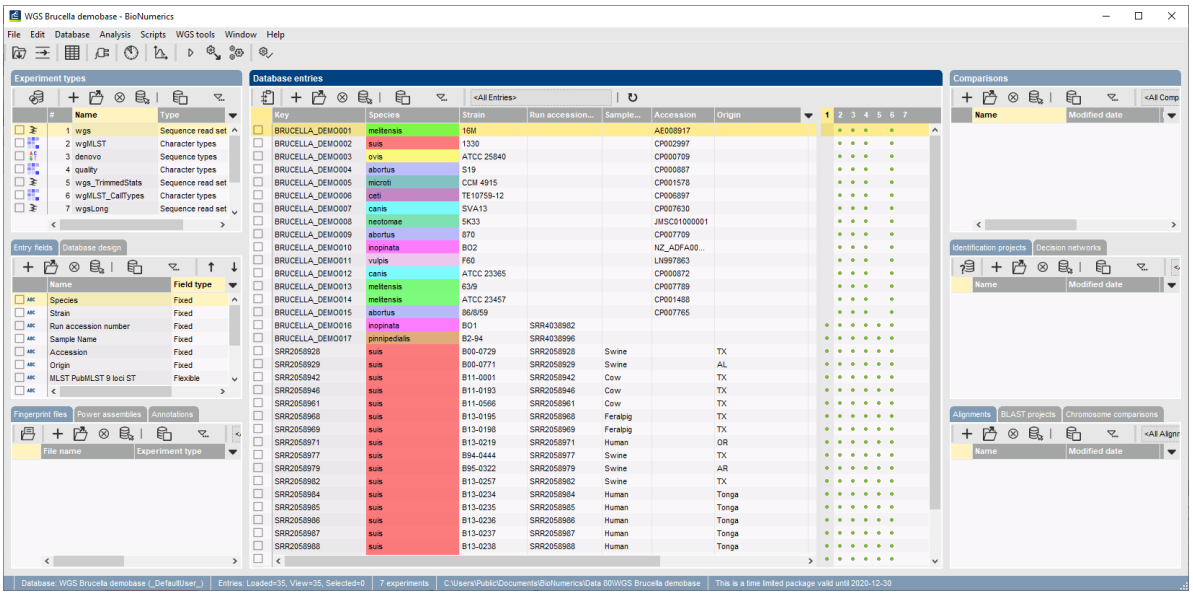

Figure 3: The *Brucella* spp. demonstration database: the *Main* window.

3 About the demonstration database

The *Brucella* spp. demonstration database contains 35 *Brucella* entries with linked data. The *WGS tools plugin* has already been installed in the demo database.

1. Select ***WGS tools*** > ***Settings...*** to access the settings of the plugin.

The calculation engine project is linked to the *Brucella* spp. allele database. No credits are assigned to this project so no jobs can be submitted to the external calculation engine, however since the option ***Enable running jobs on my own computer*** is checked in the *Calculation engine* tab, it is possible to run jobs on your own computer (see Figure 4).

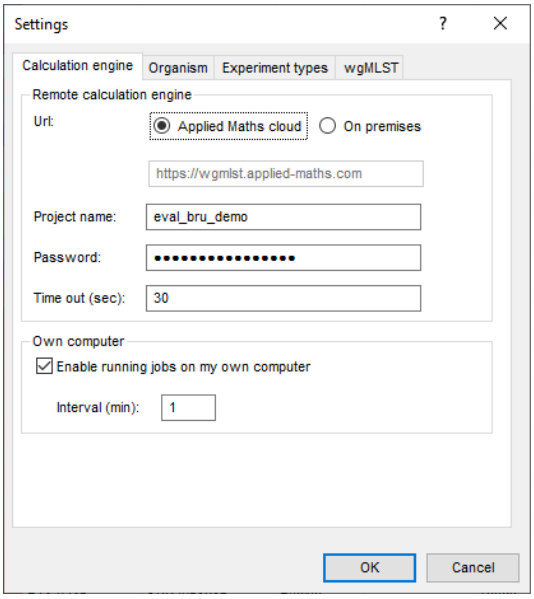

Figure 4: The *Calculation engine* tab of the *Calculation engine settings* dialog box.

2. In the *Experiment types* tab, the experiment types that are created during installation of the plugin are listed. These experiments are automatically linked to the datasets used for wgMLST analysis:

- Character experiment type **wgMLST** contains the allele calls for detected loci in each sample, where the consensus from assembly-based and assembly-free calling (if performed) resulted in a single allele ID.
- Sequence read set type **wgs** contains the link to the sequence read files on NCBI.
- Sequence experiment type **denovo** contains (1) the imported (assembled) whole genome sequences, or (2) the concatenated de novo contig sequences resulting from the de novo assembly performed on the sequence read sets.
- Character experiment type **quality** contains quality statistics for the raw data (if available), the de novo assembly (if calculated) and the allele identification algorithm(s).

No data is available for the sequence read set type **wgsLong** in the demo database. This sequence read set is used to store links to long read sequence read data (e.g. PacBio or MinION datasets).

3. Click on the *wgMLST tab* and press the <***Auto submission criteria***> button.

By default, the ***Use nomenclature acceptance criteria*** option will be checked, meaning that the automatic submission settings are defined by the curator of the allele database.

4. Click <***Cancel***> in both dialog boxes.

Additional information (in entry info fields Strain, Sample name, Origin, etc.) was collected from the corresponding publications and added to the demonstration database.

The 35 *Brucella* entries can be divided in three categories:

- Published genome assemblies are imported for 15 database entries (Key **BRUCELLA_ DEMO001** to **BRUCELLA_ DEMO015**) and 2 database entries (Key **BRUCELLA_ DEMO016** and **BRUCELLA_ DEMO017**) contain the link to sequence read set data NCBI's sequence read archive (SRA). These 17 entries represent the covered diversity of the *Brucella* spp. schema.
- 17 *Brucella suis* samples contain links to sequence read set data on NCBI's sequence read archive (SRA).
- 1 sample is present that has not yet been allocated to a species.

By clicking on one of the green dots next to an entry in the database, the corresponding results can be viewed, either in a separate window or in an experiment card for the character data types:

5. Click on a 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 5).

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

7. Click on the green colored dot for one of the entries in the second column in the *Experiment presence* panel. Column 2 corresponds to the second experiment type listed in the *Experiment types* panel, which is **wgMLST** in the default configuration.

Character experiment type **wgMLST** contains the allele calls for detected loci in each sample, where the consensus from assembly-based and assembly-free calling (if performed) resulted in a single allele ID (see Figure 6).

8. Close the character experiment card by clicking on the triangle in the top left corner.

9. 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.

Depending on the entry, the *Sequence editor* window contains (1) the imported (assembled) whole genome sequence, or (2) the concatenated de novo contig sequences resulting from the de novo assembly performed on the sequence read set.

10. Close the *Sequence editor* window.

11. Click on the green colored dot in column 4 to open the **quality** character card (default configuration) for an entry in the database.

The **quality** character card contains quality statistics for the raw data (if present), the de novo assembly (if calculated) and the different allele identification algorithm(s).

12. Close the character experiment card by clicking on the triangle in the top left corner.

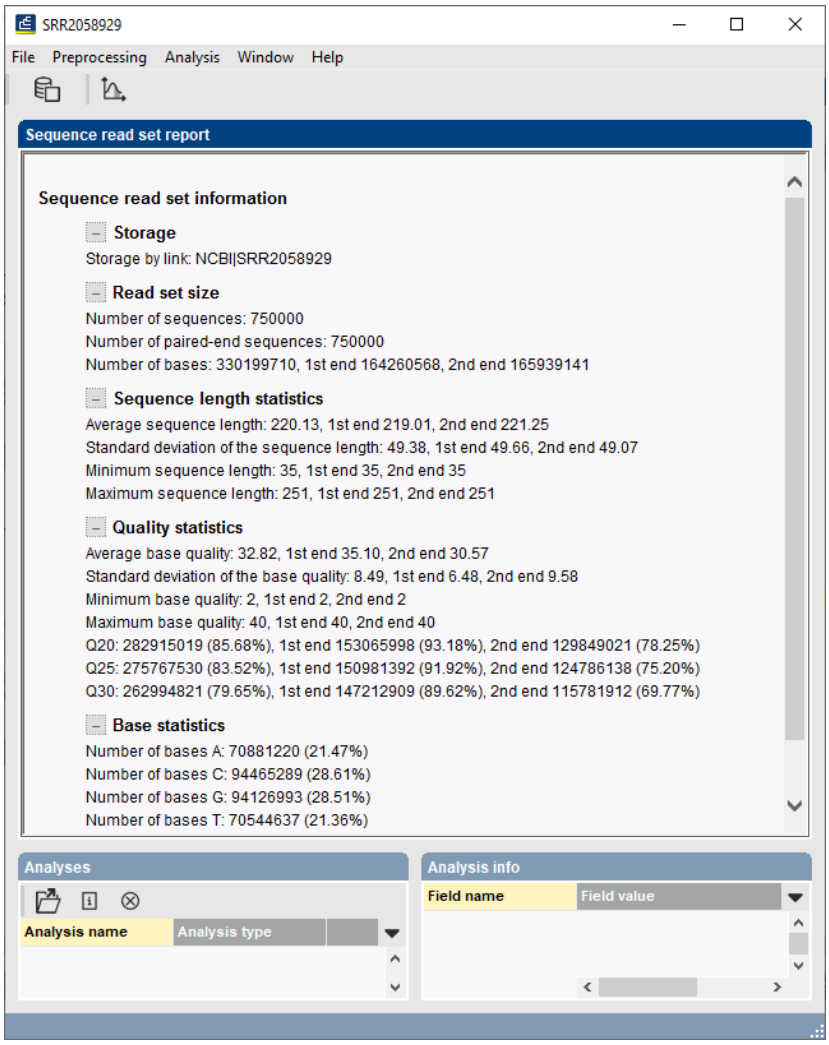

Figure 5: The *Sequence read set experiment* window.

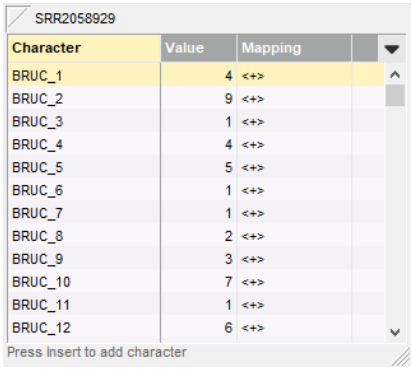

Figure 6: Character card.

4 Subschemes

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

experiment.

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

2. Click on the drop-down bar in the toolbar (see Figure 7 for an example).

The views that have been defined at the curator level are synchronized upon installation and are listed. In the *Brucella* spp. database following views are defined by the curator (see Figure 7): the default view **All loci**, the genus-wide **TRUNK** loci view, 11 species specific loci views, the MLST 9 loci view (**MLST PubMLST 9 loci**) and the MLST 21 loci view (**MLST PubMLST 21 loci**).

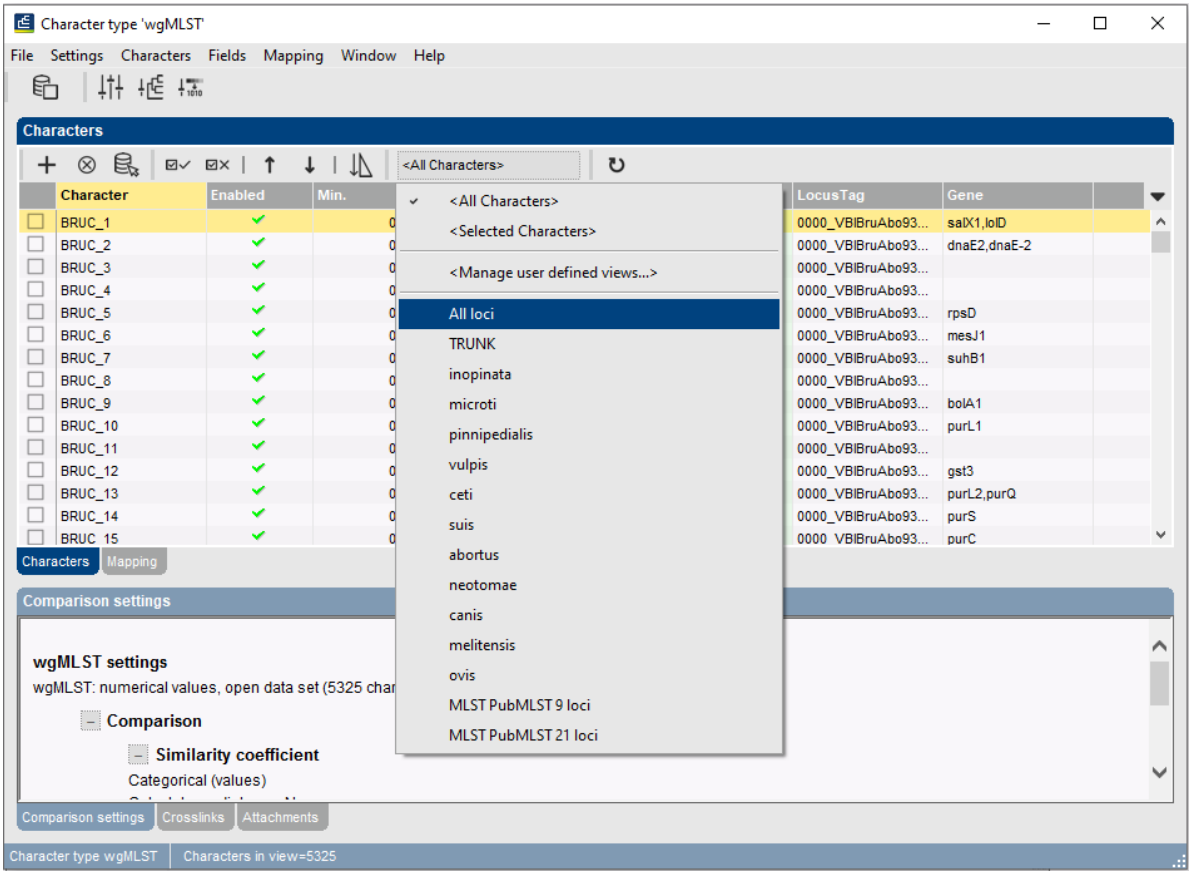

Figure 7: Character views defined by the curator.

3. 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.

4. 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 via ***Characters*** > ***Character Views*** > ***Manage user defined views...*** (<All Characters>).

5. Close the *Character type* window.

5 Sequence type assignment

Sequence types can be assigned for selected entries, based on a specific wgMLST subscheme. Note that only some of the curator-defined subschemes have associated sequence types.

1. Select the entries for which you would like to assign sequence types. For this example, select all entries with ***Edit*** > ***Select all*** (**Ctrl+A**).

2. Select ***WGS tools*** > ***Assign wgMLST sequence types...***.

This opens the *Assign sequence types* dialog box, where all available typing schemes can be checked to be included in the assignment of the sequence types (see Figure 8).

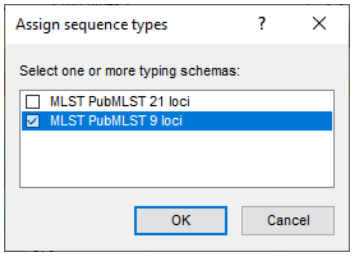

Figure 8: Sequence type assignment.

3. Only check the schema ***MLST PubMLST 9 loci*** to assign sequence types based on the 9 loci used for traditional MLST analysis and press <***OK***>.

The sequence type results are saved to an entry information field (one information field for each typing scheme). In our example, a sequence type number is added in the field **MLST PubMLST 9 loci ST** for all entries (see Figure 9).

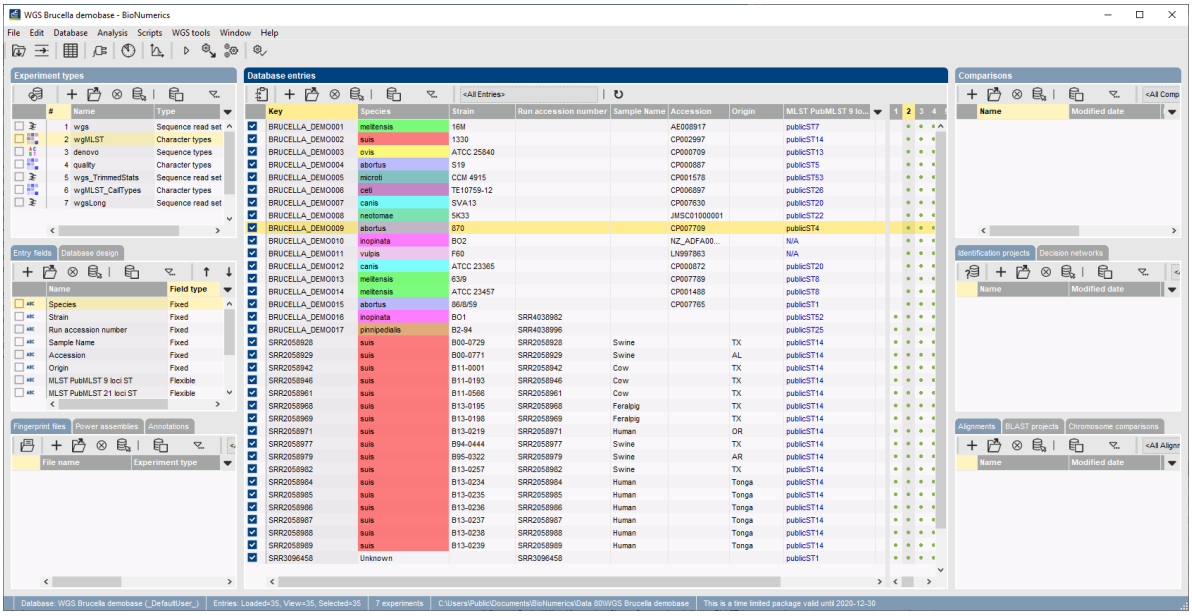

Figure 9: The *Main* window after sequence type assignment.

4. Press <***F4***> to clear the entry selection.

6 Follow-up analysis

A cluster analysis on the **wgMLST** character experiment (or a subscheme thereof) is created in the *Comparison* window or the *Advanced cluster analysis* window. The steps to create a new comparison and to perform cluster analysis on wgMLST data are explained in the next sections. The trunk-and-branch structure of the *Brucella* spp. schema allows flexibility in how to analyze different types of samples. Three possible scenarios are outlined here to illustrate this:

6.1 Compare multiple Brucella species

When you have multiple species of the genus *Brucella* in the same database and you want to compare them to each other, you will want to use the genus-wide subschema **TRUNK**.

1. In the *Database entries* panel of the *Main* window, select the entries with Key **BRUCELLA_ DEMO001** to **BRUCELLA_ DEMO017**. Selected entries are marked by a checked ballot box ( ).
2. Highlight the *Comparisons* panel in the *Main* window and select ***Edit*** > ***Create new object...*** (+) to create a new comparison for the 17 selected entries.
3. Select the **wgMLST** character experiment in the *Experiments* panel of the *Comparison* window.

Since we are dealing with multiple species, we will use the genus-wide ***TRUNK*** subschema.

4. Make sure the **TRUNK** aspect is selected for the **wgMLST** experiment (see Figure 10).
5. In the *Experiments* panel click on the eye icon ( ) that proceeds **wgMLST** to display the values of the selected aspect.

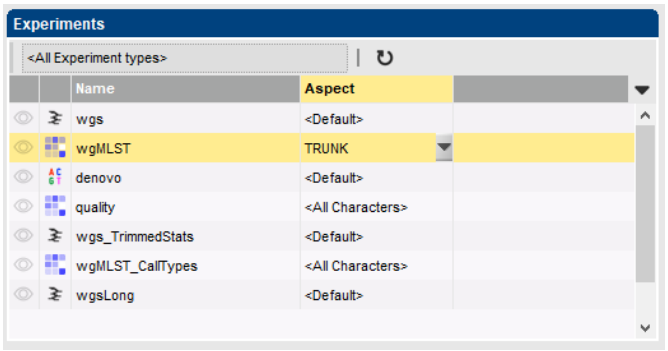

Figure 10: The TRUNK subschema.

6. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...*** and choose the ***Categorical (values)*** coefficient from the list.
7. Press <***Next***>, choose ***Complete Linkage*** in the last step and press <***Finish***>.

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 (TRUNK)**.

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 similarity values on the nodes, select ***Clustering*** > ***Dendrogram display settings...*** ( ), and tick the option ***Show node information***. Press <***OK***>.

10. Right-click on the column header of **Species** in the *Information fields* panel and select ***Create groups from database field***. In the *Group creation preferences* dialog box, leave the settings at their defaults and press <***OK***>.

11. Select ***Clustering*** > ***Dendrogram display settings...*** () again, and tick the option ***Show group colors***. Press <***OK***>.

The group colors are now displayed on the dendrogram. The *Comparison* window should now look like Figure 11.

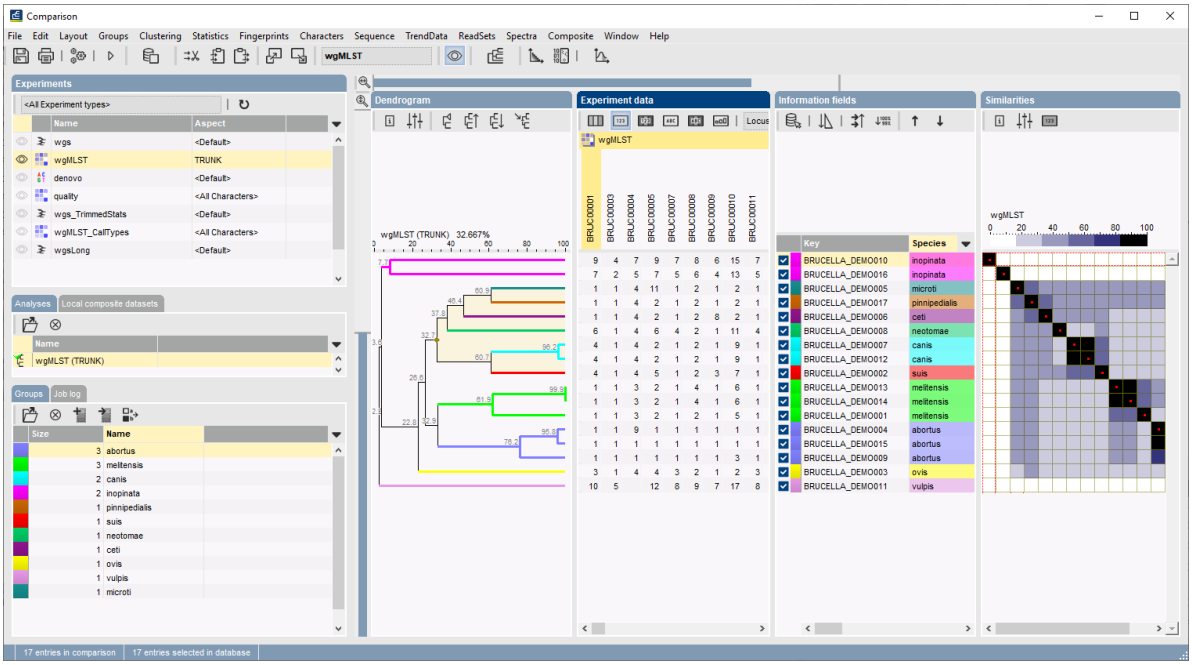

Figure 11: The *Comparison* window.

Another analysis tool that can be applied on wgMLST data is the calculation of a Minimum spanning tree (MST). A minimum spanning tree is calculated in the *Advanced cluster analysis* window which is launched from the *Comparison* window.

12. 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.

13. Specify an analysis name, make sure the **TRUNK** subscheme is selected, select ***MST for categorical data***, and press <***Next***> (see Figure 12).

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. 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.

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

15. 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”.

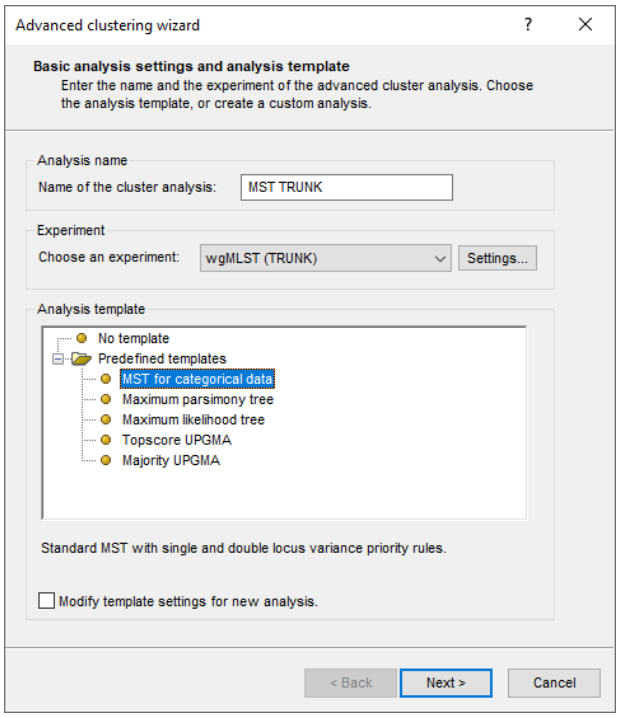

Figure 12: Settings for the calculation of the MST.

16. In the *Node labels and sizes panel*, check ***Show node labels*** and choose ***MLST PubMLST 9 loci ST*** as ***Label***.

17. Click <***OK***> to close the *Display settings* dialog box. The MST is now displayed with branch and node labels (see Figure 13).

18. Close the *Advanced cluster analysis* window.

19. Save the comparison with ***File*** > ***Save as...***. Specify a name (e.g. **Genus-wide comparison**) and close the comparison with ***File*** > ***Exit***.

6.2 Compare single Brucella species

If you have a set of samples for which you determined the species via different methods, you can use the corresponding species-specific subschema to analyze them.

20. In the *Main* window, select all *Brucella suis* entries: press <***F4***> to unselect all entries, select ***Edit*** > ***Find object in list...*** (, **Ctrl+Shift+F**), search for “suis” and select <***Select all***>.

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

22. Select the **wgMLST** character experiment in the *Experiments* panel of the *Comparison* window.

Since we are dealing with a single species, we can use the species-specific **suis** subschema.

23. Make sure the **suis** aspect is selected for the **wgMLST** experiment (see Figure 14).

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

25. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***.

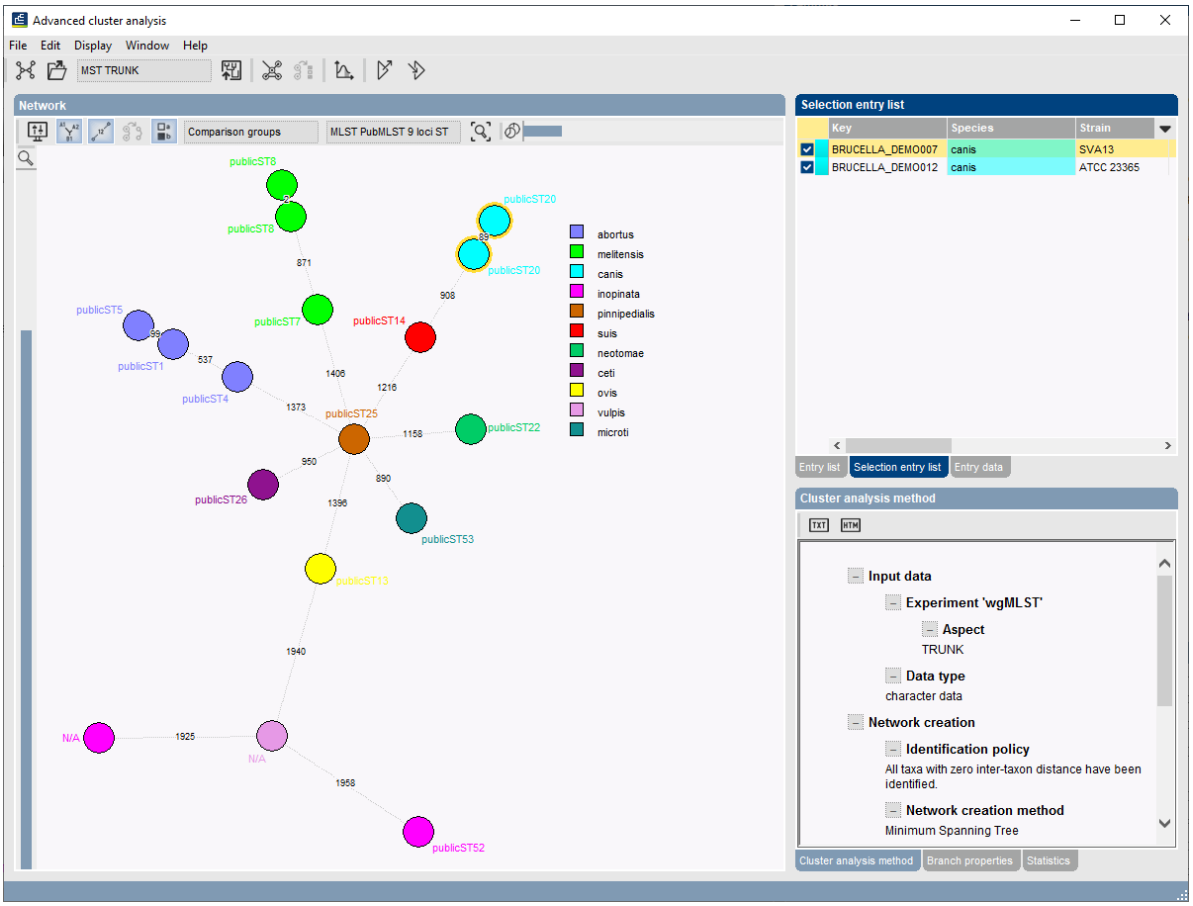

Figure 13: The minimum spanning tree based on the TRUNK schema.

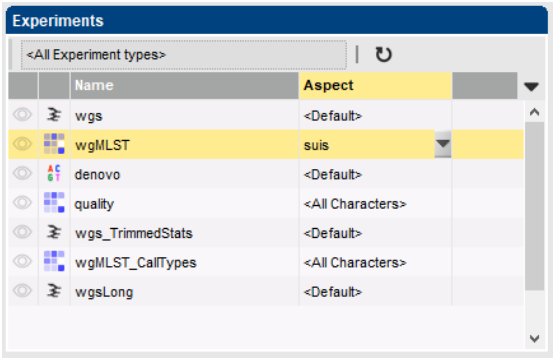

Figure 14: The **suis** subschema.

26. Since we want to compare closely related isolates choose the ***Categorical (differences)*** coefficient from the list. Specify a ***Scaling factor*** of 1.

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.

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

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 (suis)**.

28. To view the number of allele differences on the branches, select ***Clustering*** > ***Dendrogram display settings...*** (), and tick the option ***Show node information***. Press <***OK***>.

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 (1 in this example).

29. Right-click on the column header of **Sample name** in the *Information fields* panel and select ***Create groups from database field***. In the *Group creation preferences* dialog box, leave the settings at their defaults and press <***OK***>.

30. Select ***Clustering*** > ***Dendrogram display settings...*** () again, and tick the option ***Show group colors***. Press <***OK***>.

The group colors are now displayed on the dendrogram, emphasizing the clustering of the samples of human origin (see *Comparison* window).

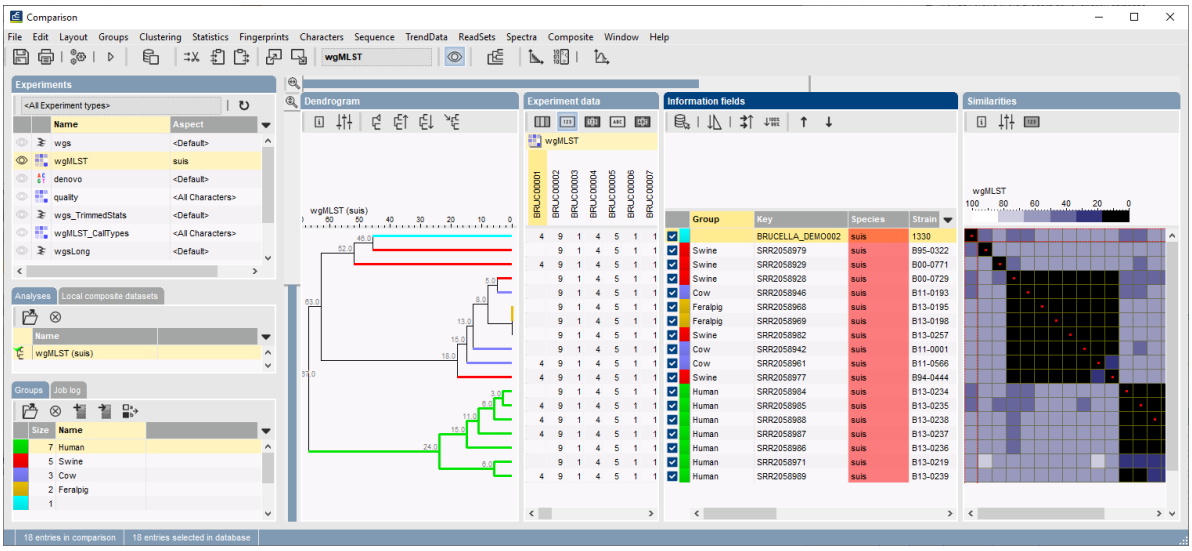

Figure 15: The *Comparison* window: clustering of species specific entries.

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

32. 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.

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

34. Select ***Clustering*** > ***Calculate*** > ***Advanced cluster analysis...*** in the *Comparison* window to launch the *Create network* wizard.

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

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

37. In the *Branch labels and sizes panel* check ***Show branch labels*** and set ***Number of digits*** to "0".

38. Click <***OK***> to close the *Display settings* dialog box. The MST is now displayed with branch labels (see Figure 16).

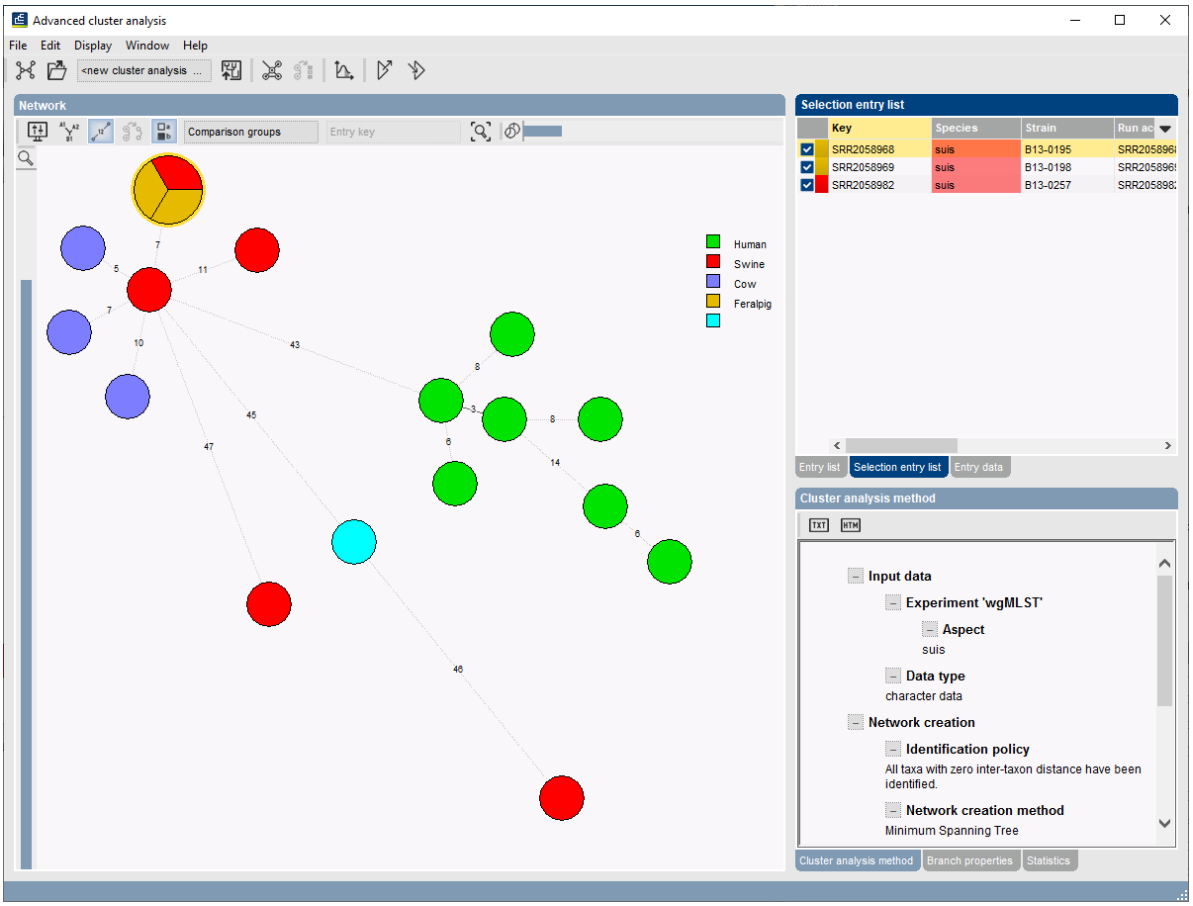

Figure 16: The minimum spanning tree based on the **suis** subschema.

39. Close the *Advanced cluster analysis* window.

40. Save the comparison with ***File*** > ***Save as...***. Specify a name (e.g. **Suis samples comparison**) and close the comparison with ***File*** > ***Exit***.

6.3 Identify unknown Brucella species

When you have a sample, from which you want to determine the type of species, you can analyze it with a species-specific subschema in the following way:

41. In the *Database entries* panel of the *Main* window press <***OK***> to clear any previous selection. Select the entries with Key **BRUCELLA_DEMO001** to **BRUCELLA_DEMO017** and include the ***Unknown*** entry.

42. Highlight the *Comparisons* panel in the *Main* window and select ***Edit*** > ***Create new object...*** (+) to create a new comparison for the 18 selected entries.

43. Select the **wgMLST** character experiment in the *Experiments* panel of the *Comparison* window.

44. To determine to which species this unknown sample relates the most, choose the **TRUNK** aspect for the **wgMLST** experiment (see Figure 10).

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

46. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...*** and choose the ***Categorical (values)*** coefficient from the list.

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

The resulting dendrogram is displayed in the *Dendrogram* panel.

The unknown sample falls within the *Brucella abortus* cluster (see Figure 17).

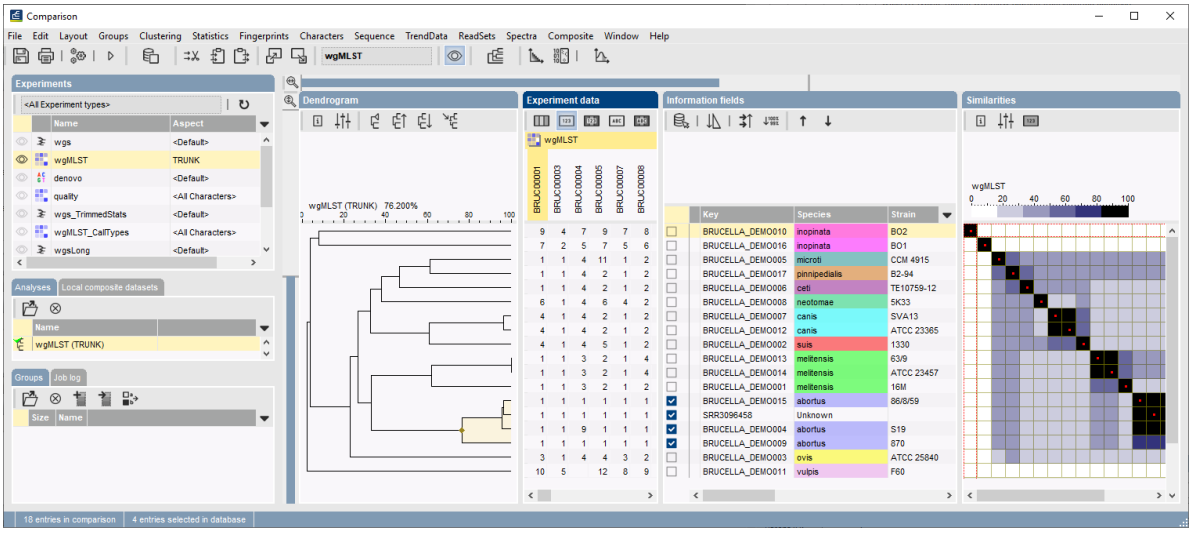

Figure 17: The *Brucella abortus* cluster.

This allows us to use the species-specific subschema.

48. Unselect all entries with <***F4***>. Select the three *B. abortus* samples and the ***Unknown*** sample.

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

50. Select the **wgMLST** character experiment in the *Experiments* panel of the *Comparison* window.

51. Make sure the **abortus** aspect is selected for the **wgMLST** experiment (see Figure 18).

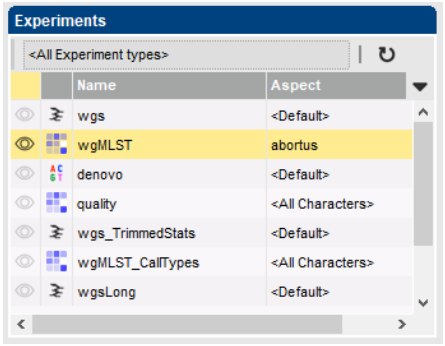

Figure 18: The **abortus** aspect.

52. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***, choose the ***Categorical (differences)*** coefficient from the list. Specify a ***Scaling factor*** of 10.

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

54. To view the number of allele differences on the branches, select ***Clustering*** > ***Dendrogram display settings...*** (), and tick the option ***Show node information***. Press <***OK***>.

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 in this example).

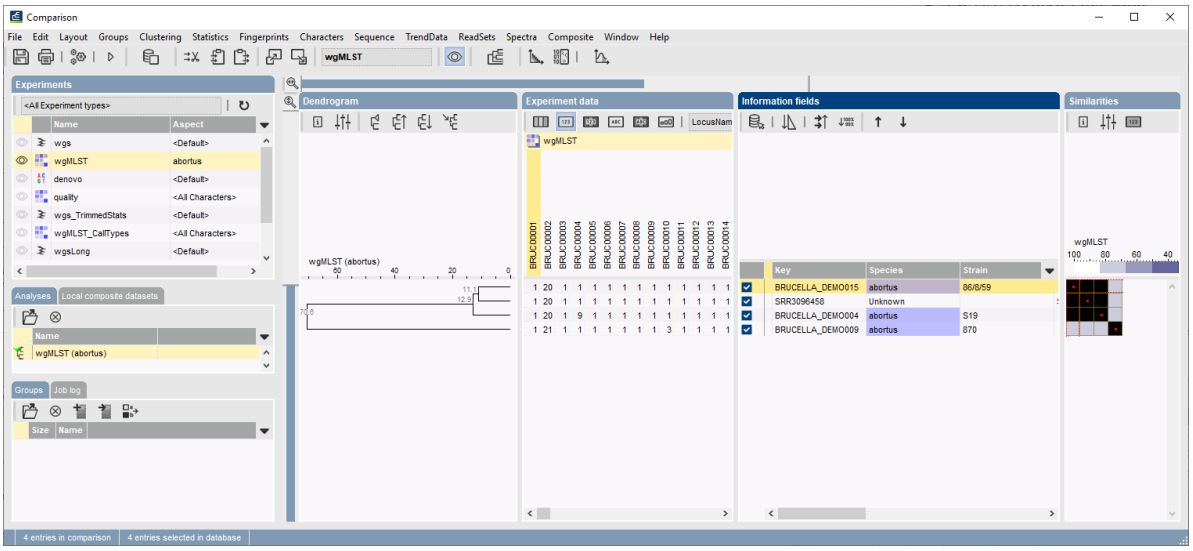

Figure 19: Clustering based on the **abortus** aspect.

55. Close the *Comparison* window.