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 create their own projects, or buy Calculation Engine credits. 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 ”Whole genome MLST typing in BioNumerics: routine workflow”
- Tutorial ”Whole genome MLST typing in BioNumerics: detailed exploration of results”
- *WGS tools plugin* manual

Furthermore, a leaflet on our website (`http://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 ***Download example databases*** link, located in the lower right corner of the *BioNumerics Startup* window.

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

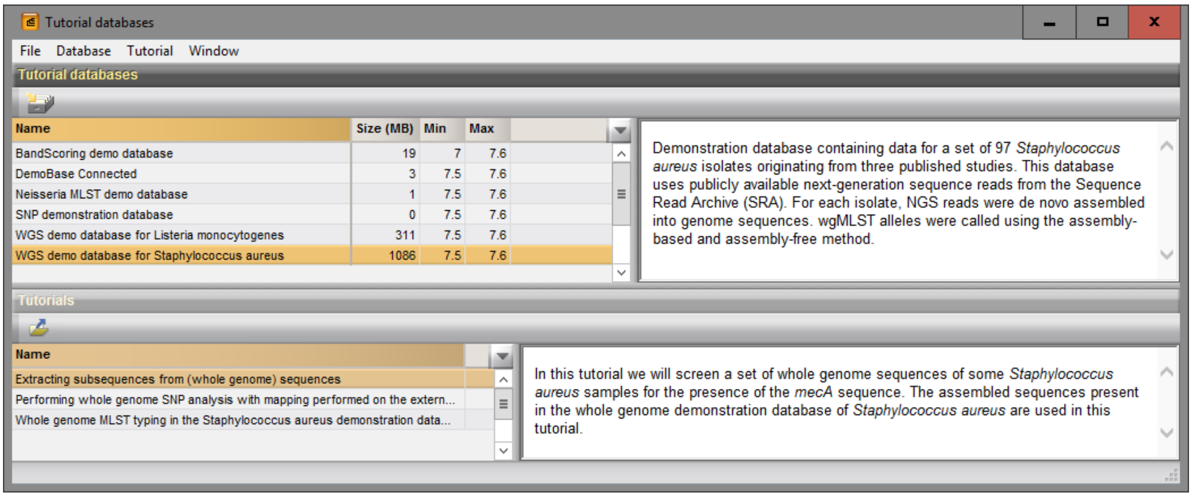
2. Select the **WGS demo database for Brucella** from the list and select ***Database*** > ***Download*** (📥).

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** appears in the *BioNumerics Startup* window.

5. Double-click the **WGS demo database for Brucella** in the *BioNumerics Startup* window to open the database.



Name	Size (MB)	Min	Max
BandScoring demo database	19	7	7.6
DemoBase Connected	3	7.5	7.6
Neisseria MLST demo database	1	7.5	7.6
SNP demonstration database	0	7.5	7.6
WGS demo database for Listeria monocytogenes	311	7.5	7.6
WGS demo database for Staphylococcus aureus	1086	7.5	7.6

Demonstration database containing data for a set of 97 *Staphylococcus aureus* isolates originating from three published studies. This database uses publicly available next-generation sequence reads from the Sequence Read Archive (SRA). For each isolate, NGS reads were de novo assembled into genome sequences. wgMLST alleles were called using the assembly-based and assembly-free method.

Name
Extracting subsequences from (whole genome) sequences
Performing whole genome SNP analysis with mapping performed on the extern...
Whole genome MLST typing in the Staphylococcus aureus demonstration data...

In this tutorial we will screen a set of whole genome sequences of some *Staphylococcus aureus* samples for the presence of the *mecA* sequence. The assembled sequences present in the whole genome demonstration database of *Staphylococcus aureus* are used in this tutorial.

Figure 1: The *Tutorial databases* window, used to download the demonstration 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 `http://www.applied-maths.com/download/sample-data`, under 'WGS demo database for Brucella'.

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*** remains 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.

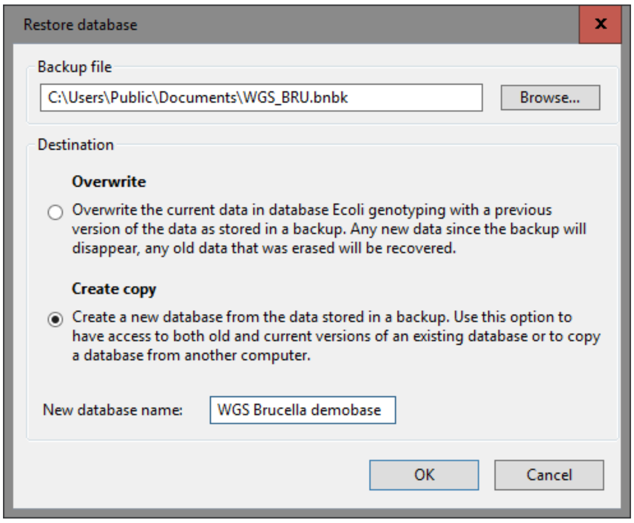
12. Once the process is complete, click <***Yes***> to open the database.

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

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.



Backup file

C:\Users\Public\Documents\WGS_BRU.bnbk

Destination

Overwrite

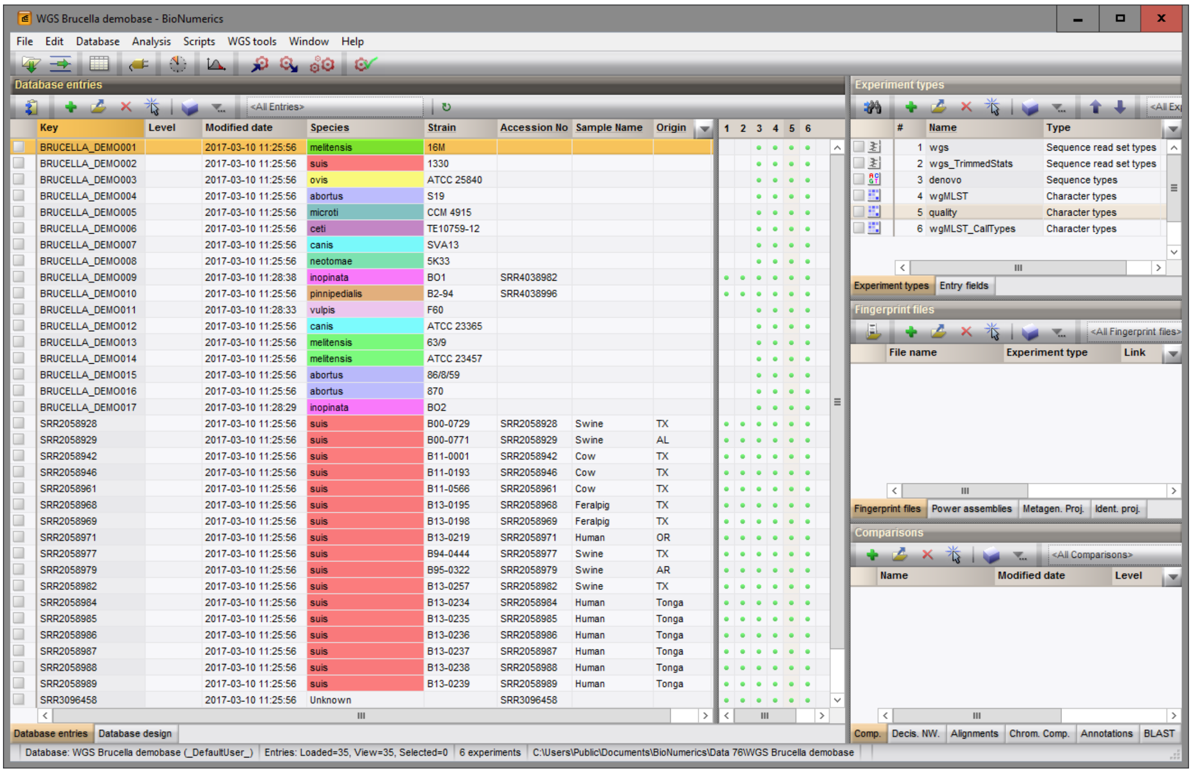
○ Overwrite the current data in database Ecoli genotyping with a previous version of the data as stored in a backup. Any new data since the backup will disappear, any old data that was erased will be recovered.

Create copy

● Create a new database from the data stored in a backup. Use this option to have access to both old and current versions of an existing database or to copy a database from another computer.

New database name: WGS Brucella demobase

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



Key	Level	Modified date	Species	Strain	Accession No	Sample Name	Origin	1	2	3	4	5	6
BRUCELLA_DEMO001		2017-03-10 11:25:56	melitensis	16M						●	●	●	●
BRUCELLA_DEMO002		2017-03-10 11:25:56	suis	1330						●	●	●	●
BRUCELLA_DEMO003		2017-03-10 11:25:56	ovis	ATCC 25840						●	●	●	●
BRUCELLA_DEMO004		2017-03-10 11:25:56	abortus	S19						●	●	●	●
BRUCELLA_DEMO005		2017-03-10 11:25:56	microti	CCM 4915						●	●	●	●
BRUCELLA_DEMO006		2017-03-10 11:25:56	ceti	TE10759-12						●	●	●	●
BRUCELLA_DEMO007		2017-03-10 11:25:56	canis	SVA13						●	●	●	●
BRUCELLA_DEMO008		2017-03-10 11:25:56	neotomae	5K33						●	●	●	●
BRUCELLA_DEMO009		2017-03-10 11:28:38	inopinata	BO1	SRR4038982			●	●	●	●	●	●
BRUCELLA_DEMO010		2017-03-10 11:25:56	pinnipedialis	B2-94	SRR4038996			●	●	●	●	●	●
BRUCELLA_DEMO011		2017-03-10 11:28:33	vulpis	F60						●	●	●	●
BRUCELLA_DEMO012		2017-03-10 11:25:56	canis	ATCC 23365						●	●	●	●
BRUCELLA_DEMO013		2017-03-10 11:25:56	melitensis	63/9						●	●	●	●
BRUCELLA_DEMO014		2017-03-10 11:25:56	melitensis	ATCC 23457						●	●	●	●
BRUCELLA_DEMO015		2017-03-10 11:25:56	abortus	86/8/59						●	●	●	●
BRUCELLA_DEMO016		2017-03-10 11:25:56	abortus	870						●	●	●	●
BRUCELLA_DEMO017		2017-03-10 11:28:29	inopinata	BO2						●	●	●	●
SRR2058928		2017-03-10 11:25:56	suis	B00-0729	SRR2058928	Swine	TX	●	●	●	●	●	●
SRR2058929		2017-03-10 11:25:56	suis	B00-0771	SRR2058929	Swine	AL	●	●	●	●	●	●
SRR2058942		2017-03-10 11:25:56	suis	B11-0001	SRR2058942	Cow	TX	●	●	●	●	●	●
SRR2058946		2017-03-10 11:25:56	suis	B11-0193	SRR2058946	Cow	TX	●	●	●	●	●	●
SRR2058961		2017-03-10 11:25:56	suis	B11-0566	SRR2058961	Cow	TX	●	●	●	●	●	●
SRR2058968		2017-03-10 11:25:56	suis	B13-0195	SRR2058968	Feralpig	TX	●	●	●	●	●	●
SRR2058969		2017-03-10 11:25:56	suis	B13-0198	SRR2058969	Feralpig	TX	●	●	●	●	●	●
SRR2058971		2017-03-10 11:25:56	suis	B13-0219	SRR2058971	Human	OR	●	●	●	●	●	●
SRR2058977		2017-03-10 11:25:56	suis	B94-0444	SRR2058977	Swine	TX	●	●	●	●	●	●
SRR2058979		2017-03-10 11:25:56	suis	B95-0322	SRR2058979	Swine	AR	●	●	●	●	●	●
SRR2058982		2017-03-10 11:25:56	suis	B13-0257	SRR2058982	Swine	TX	●	●	●	●	●	●
SRR2058984		2017-03-10 11:25:56	suis	B13-0234	SRR2058984	Human	Tonga	●	●	●	●	●	●
SRR2058985		2017-03-10 11:25:56	suis	B13-0235	SRR2058985	Human	Tonga	●	●	●	●	●	●
SRR2058986		2017-03-10 11:25:56	suis	B13-0236	SRR2058986	Human	Tonga	●	●	●	●	●	●
SRR2058987		2017-03-10 11:25:56	suis	B13-0237	SRR2058987	Human	Tonga	●	●	●	●	●	●
SRR2058988		2017-03-10 11:25:56	suis	B13-0238	SRR2058988	Human	Tonga	●	●	●	●	●	●
SRR2058989		2017-03-10 11:25:56	suis	B13-0239	SRR2058989	Human	Tonga	●	●	●	●	●	●
SRR3096458		2017-03-10 11:25:56	Unknown		SRR3096458			●	●	●	●	●	●

#	Name	Type
1	wgs	Sequence read set types
2	wgs_TrimmedStats	Sequence read set types
3	denovo	Sequence types
4	wgMLST	Character types
5	quality	Character types
6	wgMLST_CallTypes	Character types

Database: WGS Brucella demobase (_DefaultUser_) | Entries: Loaded=35, View=35, Selected=0 | 6 experiments | C:\Users\Public\Documents\BioNumerics\Data 76\WGS Brucella demobase

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

The calculation engine project is linked to the *Brucella* spp. allele database. Note that no credits are assigned to this project so no jobs can be submitted.

2. In the *Experiment types* tab, the four 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).

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 4).

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.

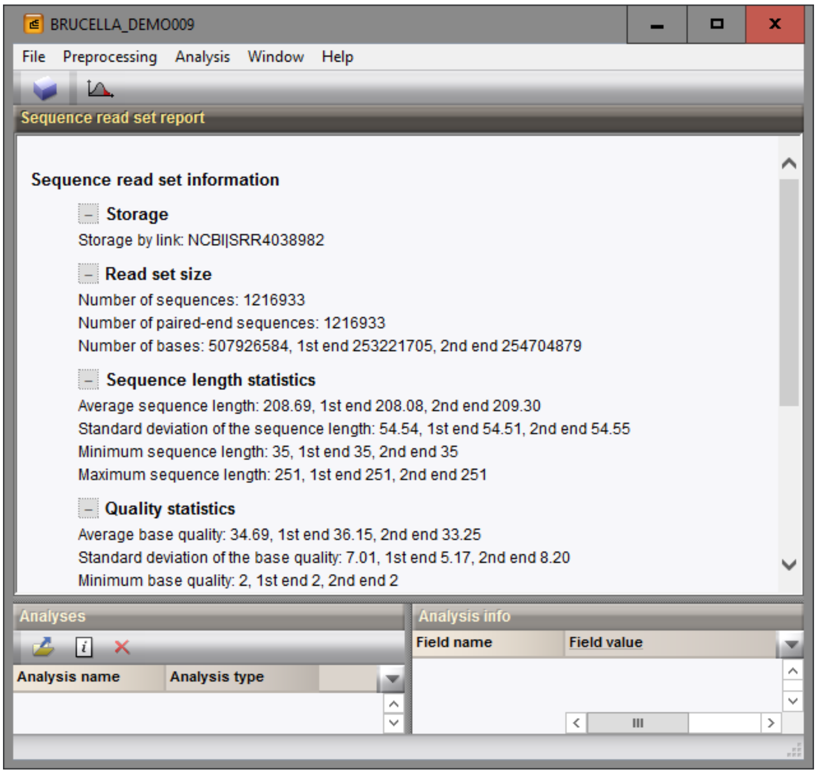
Character experiment type **wgMLST** contains the allele calls for detected loci in each sample, where the consensus from assembly-based and assembly-free calling (if performed) resulted in a single allele ID (see Figure 5).

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.



Sequence read set information

Storage

Storage by link: NCBI|SRR4038982

Read set size

Number of sequences: 1216933
Number of paired-end sequences: 1216933
Number of bases: 507926584, 1st end 253221705, 2nd end 254704879

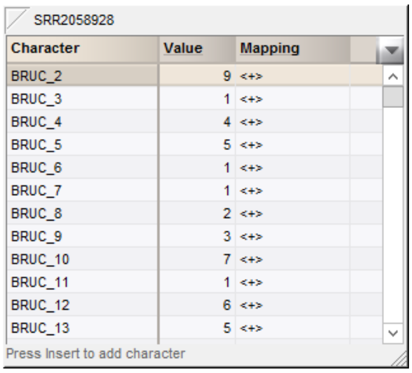
Sequence length statistics

Average sequence length: 208.69, 1st end 208.08, 2nd end 209.30
Standard deviation of the sequence length: 54.54, 1st end 54.51, 2nd end 54.55
Minimum sequence length: 35, 1st end 35, 2nd end 35
Maximum sequence length: 251, 1st end 251, 2nd end 251

Quality statistics

Average base quality: 34.69, 1st end 36.15, 2nd end 33.25
Standard deviation of the base quality: 7.01, 1st end 5.17, 2nd end 8.20
Minimum base quality: 2, 1st end 2, 2nd end 2

Figure 4: The *Sequence read set experiment* window.


SRR2058928

Character	Value	Mapping
BRUC_2	9	<+>
BRUC_3	1	<+>
BRUC_4	4	<+>
BRUC_5	5	<+>
BRUC_6	1	<+>
BRUC_7	1	<+>
BRUC_8	2	<+>
BRUC_9	3	<+>
BRUC_10	7	<+>
BRUC_11	1	<+>
BRUC_12	6	<+>
BRUC_13	5	<+>

Press Insert to add character

Figure 5: Character card.

11. Click on the green colored dot in column 4 to open the **quality** character card 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.

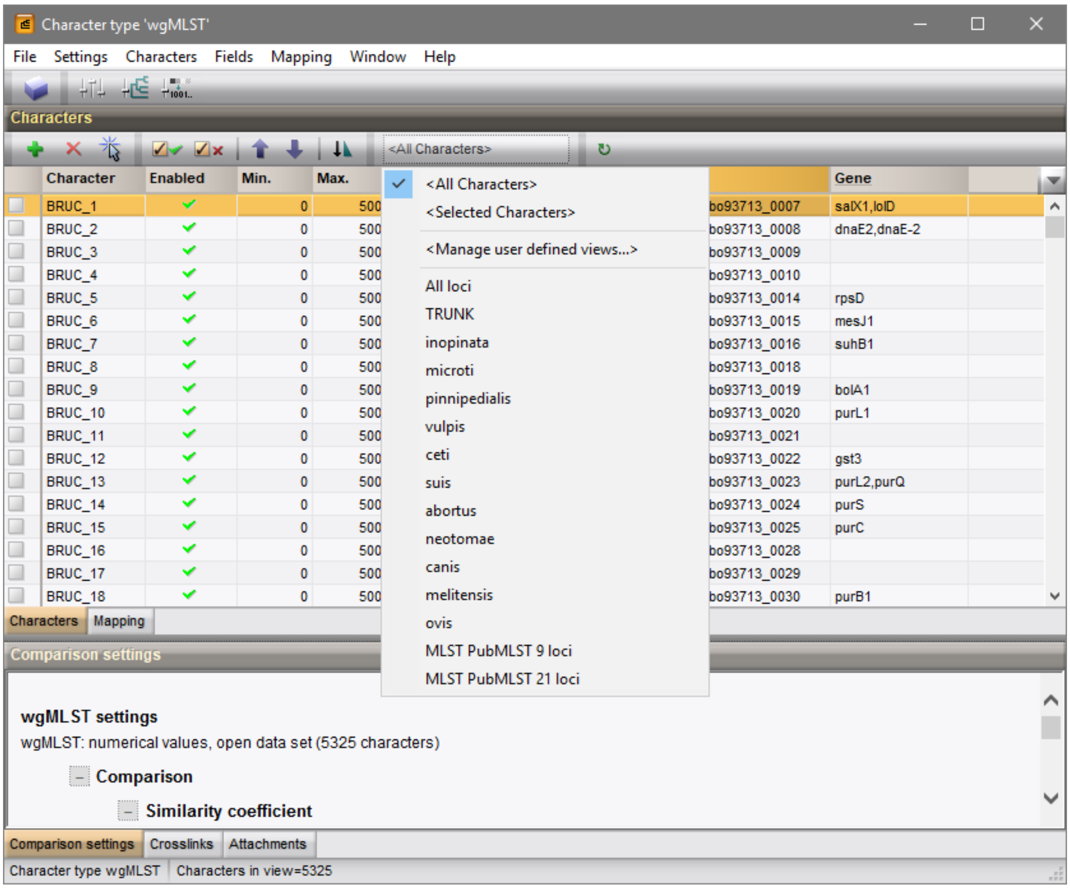
4 Subschemes

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 6 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 6): 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**).



Character	Enabled	Min.	Max.		Gene
BRUC_1	✓	0	500	bo93713_0007	salX1,lolD
BRUC_2	✓	0	500	bo93713_0008	dnaE2,dnaE-2
BRUC_3	✓	0	500	bo93713_0009	
BRUC_4	✓	0	500	bo93713_0010	
BRUC_5	✓	0	500	bo93713_0014	rpsD
BRUC_6	✓	0	500	bo93713_0015	mesJ1
BRUC_7	✓	0	500	bo93713_0016	suhB1
BRUC_8	✓	0	500	bo93713_0018	
BRUC_9	✓	0	500	bo93713_0019	bolA1
BRUC_10	✓	0	500	bo93713_0020	purL1
BRUC_11	✓	0	500	bo93713_0021	
BRUC_12	✓	0	500	bo93713_0022	gst3
BRUC_13	✓	0	500	bo93713_0023	purL2,purQ
BRUC_14	✓	0	500	bo93713_0024	purS
BRUC_15	✓	0	500	bo93713_0025	purC
BRUC_16	✓	0	500	bo93713_0028	
BRUC_17	✓	0	500	bo93713_0029	
BRUC_18	✓	0	500	bo93713_0030	purB1

wgMLST settings

wgMLST: numerical values, open data set (5325 characters)

- Comparison
  - Similarity coefficient

Character type wgMLST | Characters in view=5325

Figure 6: 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.

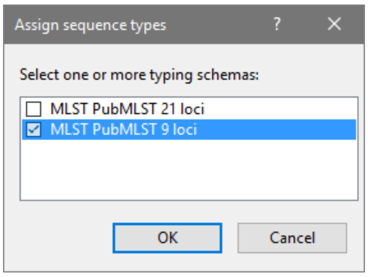
5 Sequence type assignment

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 7).



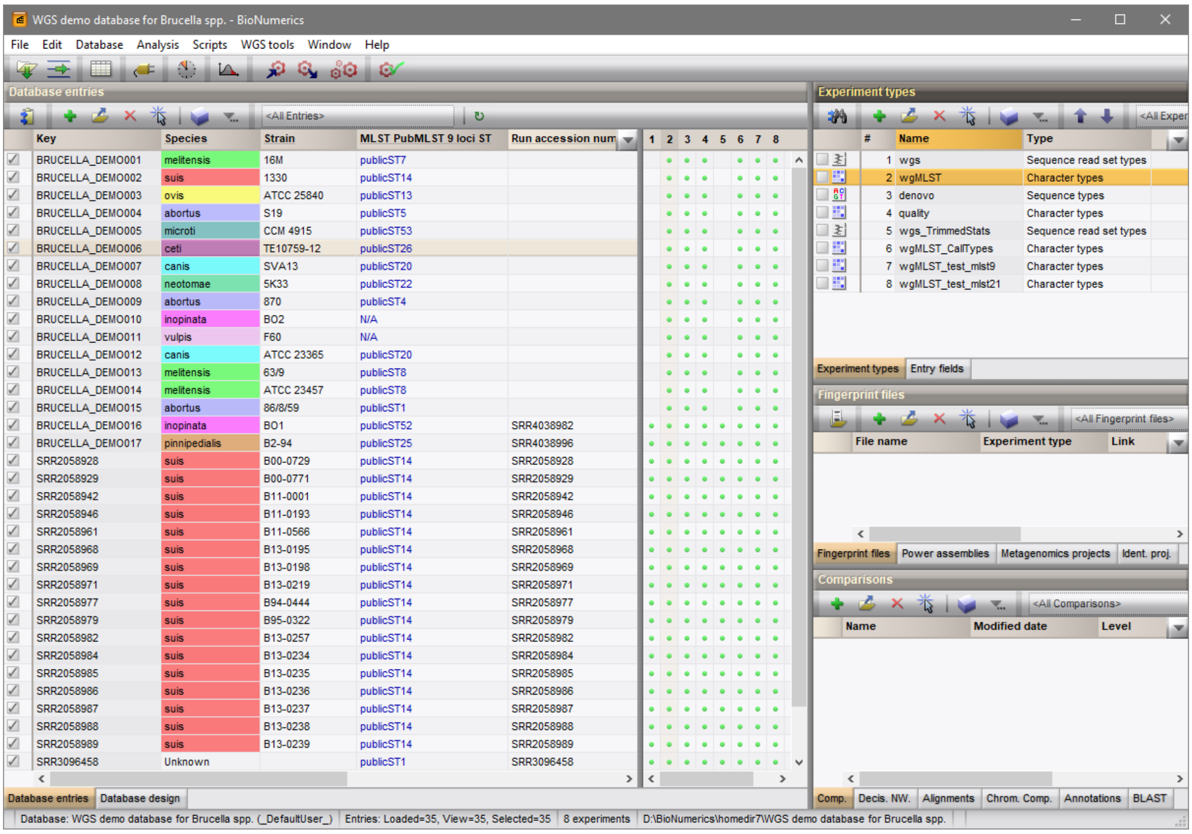
Select one or more typing schemas:

☐ MLST PubMLST 21 loci
☑ MLST PubMLST 9 loci

Figure 7: 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 8).



	Key	Species	Strain	MLST PubMLST 9 loci ST	Run accession num
☑	BRUCELLA_DEMO001	melitensis	16M	publicST7	
☑	BRUCELLA_DEMO002	suis	1330	publicST14	
☑	BRUCELLA_DEMO003	ovis	ATCC 25840	publicST13	
☑	BRUCELLA_DEMO004	abortus	S19	publicST5	
☑	BRUCELLA_DEMO005	microti	CCM 4915	publicST53	
☑	BRUCELLA_DEMO006	ceti	TE10759-12	publicST26	
☑	BRUCELLA_DEMO007	canis	SVA13	publicST20	
☑	BRUCELLA_DEMO008	neotomae	5K33	publicST22	
☑	BRUCELLA_DEMO009	abortus	870	publicST4	
☑	BRUCELLA_DEMO010	inopinata	BO2	N/A	
☑	BRUCELLA_DEMO011	vulpis	F60	N/A	
☑	BRUCELLA_DEMO012	canis	ATCC 23365	publicST20	
☑	BRUCELLA_DEMO013	melitensis	63/9	publicST8	
☑	BRUCELLA_DEMO014	melitensis	ATCC 23457	publicST8	
☑	BRUCELLA_DEMO015	abortus	86/8/59	publicST1	
☑	BRUCELLA_DEMO016	inopinata	BO1	publicST52	SRR4038982
☑	BRUCELLA_DEMO017	pinnipedialis	B2-94	publicST25	SRR4038996
☑	SRR2058928	suis	B00-0729	publicST14	SRR2058928
☑	SRR2058929	suis	B00-0771	publicST14	SRR2058929
☑	SRR2058942	suis	B11-0001	publicST14	SRR2058942
☑	SRR2058946	suis	B11-0193	publicST14	SRR2058946
☑	SRR2058961	suis	B11-0566	publicST14	SRR2058961
☑	SRR2058968	suis	B13-0195	publicST14	SRR2058968
☑	SRR2058969	suis	B13-0198	publicST14	SRR2058969
☑	SRR2058971	suis	B13-0219	publicST14	SRR2058971
☑	SRR2058977	suis	B94-0444	publicST14	SRR2058977
☑	SRR2058979	suis	B95-0322	publicST14	SRR2058979
☑	SRR2058982	suis	B13-0257	publicST14	SRR2058982
☑	SRR2058984	suis	B13-0234	publicST14	SRR2058984
☑	SRR2058985	suis	B13-0235	publicST14	SRR2058985
☑	SRR2058986	suis	B13-0236	publicST14	SRR2058986
☑	SRR2058987	suis	B13-0237	publicST14	SRR2058987
☑	SRR2058988	suis	B13-0238	publicST14	SRR2058988
☑	SRR2058989	suis	B13-0239	publicST14	SRR2058989
☑	SRR3096458	Unknown		publicST1	SRR3096458

	#	Name	Type
☐	1	wgs	Sequence read set types
☐	2	wgMLST	Character types
☐	3	denovo	Sequence types
☐	4	quality	Character types
☐	5	wgs_TrimmedStats	Sequence read set types
☐	6	wgMLST_CallTypes	Character types
☐	7	wgMLST_test_mlst9	Character types
☐	8	wgMLST_test_mlst21	Character types

Database: WGS demo database for Brucella spp. (_DefaultUser_) | Entries: Loaded=35, View=35, Selected=35 | 8 experiments | D:\BioNumerics\homedir7\WGS demo database for Brucella spp.

Figure 8: The *Main* window after sequence type assignment.

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:

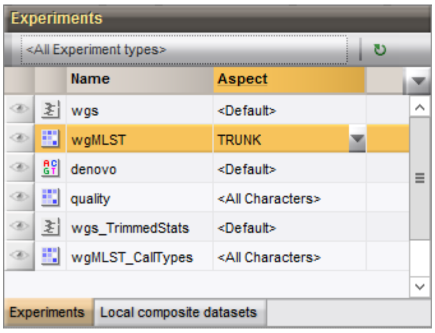
6.1 Compare multiple Brucella species

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**.
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 9).
5. In the *Experiments* panel click on the eye icon (👁) that proceeds **wgMLST** to display the values of the selected aspect.



Name	Aspect
wgs	<Default>
wgMLST	TRUNK
denovo	<Default>
quality	<All Characters>
wgs_TrimmedStats	<Default>
wgMLST_CallTypes	<All Characters>

Figure 9: The TRUNK subschema.

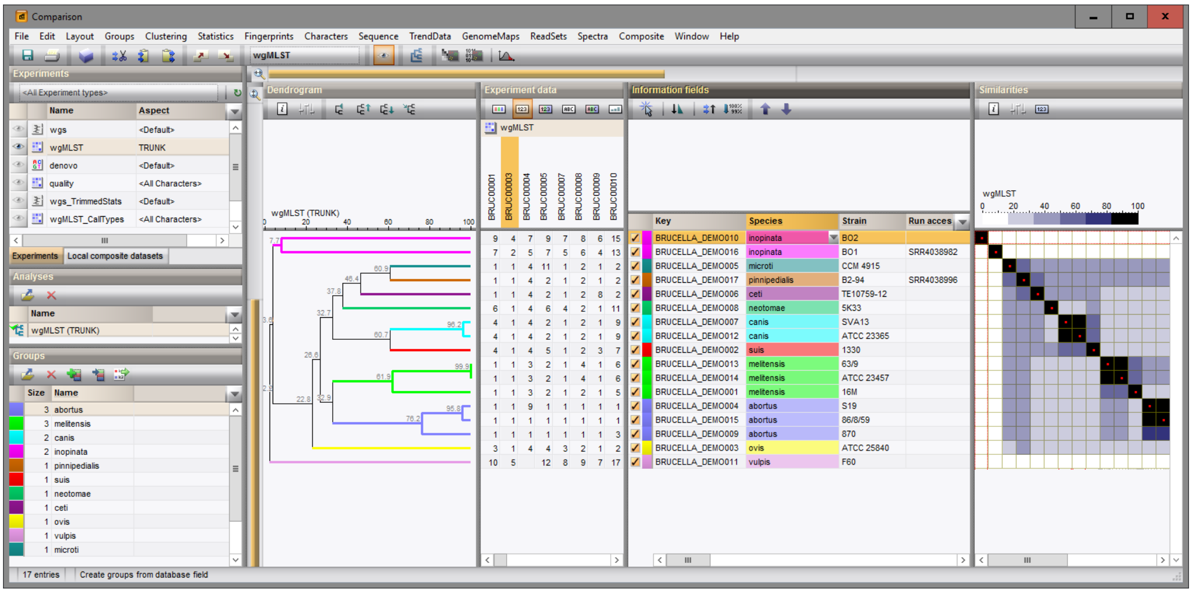
6. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...*** and choose the ***Categorical (values)*** coefficient from the list.
7. Press <***Next***>, choose ***Complete Linkage*** in the last step and press <***Finish***>.

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*** (ℹ).
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 10.



Name	Aspect
wgs	<Default>
wgMLST	TRUNK
denovo	<Default>
quality	<All Characters>
wgs_TrimmedStats	<Default>
wgMLST_CallTypes	<All Characters>

Name
wgMLST (TRUNK)

	Size	Name
🟦	3	abortus
🟩	3	melitensis
🟦	2	canis
🟪	2	inopinata
🟧	1	pinnipedialis
🟥	1	suis
🟩	1	neotomae
🟪	1	ceti
🟨	1	ovis
🟪	1	vulpis
🟦	1	microti

17 entries | Create groups from database field

BRUC00001	BRUC00003	BRUC00004	BRUC00005	BRUC00007	BRUC00008	BRUC00009	BRUC00010		Key	Species	Strain	Run acces
9	4	7	9	7	8	6	15	✓	BRUCELLA_DEMO010	inopinata	BO2	
7	2	5	7	5	6	4	13	✓	BRUCELLA_DEMO016	inopinata	BO1	SRR4038982
1	1	4	11	1	2	1	2	✓	BRUCELLA_DEMO005	microti	CCM 4915	
1	1	4	2	1	2	1	2	✓	BRUCELLA_DEMO017	pinnipedialis	B2-94	SRR4038996
1	1	4	2	1	2	8	2	✓	BRUCELLA_DEMO006	ceti	TE10759-12	
6	1	4	6	4	2	1	11	✓	BRUCELLA_DEMO008	neotomae	5K33	
4	1	4	2	1	2	1	9	✓	BRUCELLA_DEMO007	canis	SVA13	
4	1	4	2	1	2	1	9	✓	BRUCELLA_DEMO012	canis	ATCC 23365	
4	1	4	5	1	2	3	7	✓	BRUCELLA_DEMO002	suis	1330	
1	1	3	2	1	4	1	6	✓	BRUCELLA_DEMO013	melitensis	63/9	
1	1	3	2	1	4	1	6	✓	BRUCELLA_DEMO014	melitensis	ATCC 23457	
1	1	3	2	1	2	1	5	✓	BRUCELLA_DEMO001	melitensis	16M	
1	1	9	1	1	1	1	1	✓	BRUCELLA_DEMO004	abortus	S19	
1	1	1	1	1	1	1	1	✓	BRUCELLA_DEMO015	abortus	86/8/59	
1	1	1	1	1	1	1	3	✓	BRUCELLA_DEMO009	abortus	870	
3	1	4	4	3	2	1	2	✓	BRUCELLA_DEMO003	ovis	ATCC 25840	
10	5		12	8	9	7	17	✓	BRUCELLA_DEMO011	vulpis	F60	

Figure 10: 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 11).

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.

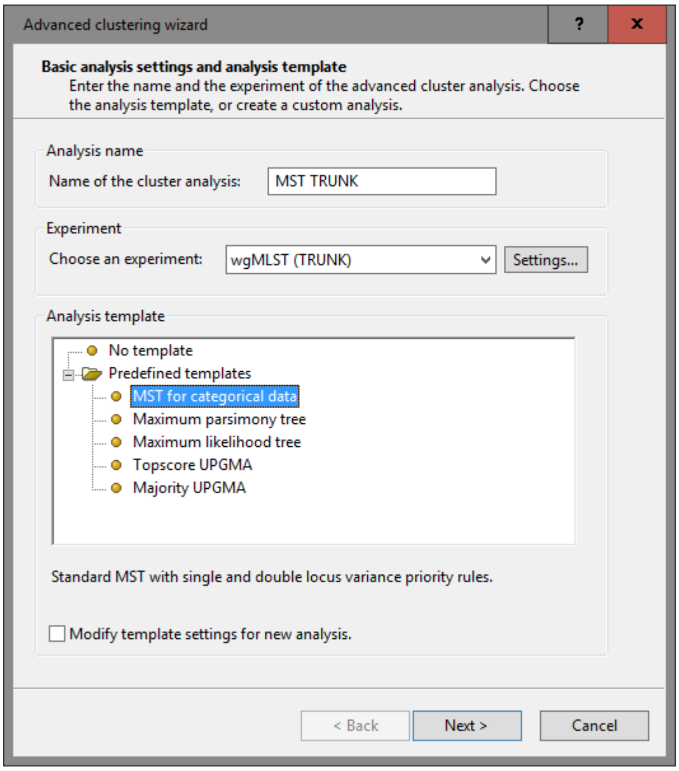
14. Press ▦ or choose ***Display*** > ***Display settings*** to open the *Display settings* dialog box.

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”.

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 12).

18. Close the *Advanced cluster analysis* window.



Basic analysis settings and analysis template
Enter the name and the experiment of the advanced cluster analysis. Choose the analysis template, or create a custom analysis.

Analysis name
Name of the cluster analysis: MST TRUNK

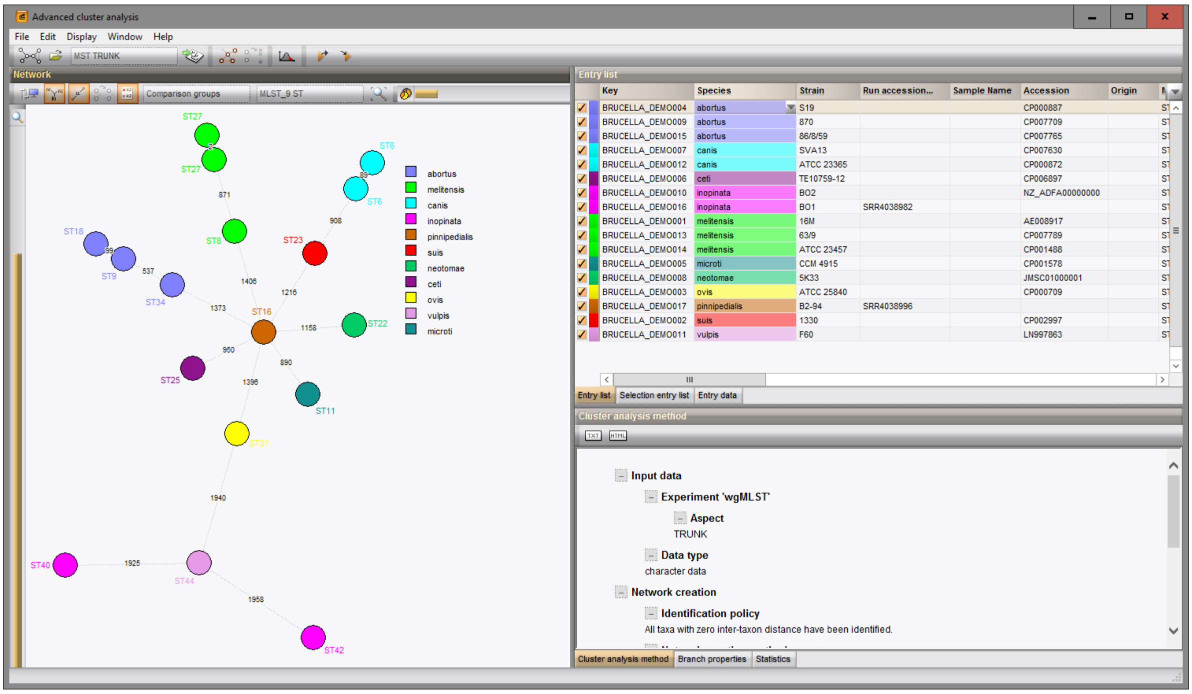
Experiment
Choose an experiment: wgMLST (TRUNK)

Analysis template
- No template
- Predefined templates
  - MST for categorical data
  - Maximum parsimony tree
  - Maximum likelihood tree
  - Topscore UPGMA
  - Majority UPGMA

Standard MST with single and double locus variance priority rules.

☐ Modify template settings for new analysis.

Figure 11: Settings for the calculation of the MST.



Key	Species	Strain	Run accession...	Sample Name	Accession	Origin
BRUCELLA_DEMO004	abortus	S19			CP000887	
BRUCELLA_DEMO009	abortus	870			CP007709	
BRUCELLA_DEMO015	abortus	86/8/59			CP007765	
BRUCELLA_DEMO007	canis	SVA13			CP007630	
BRUCELLA_DEMO012	canis	ATCC 23365			CP000872	
BRUCELLA_DEMO006	ceti	TE10759-12			CP006897	
BRUCELLA_DEMO010	inopinata	BO2			NZ_ADFA00000000	
BRUCELLA_DEMO016	inopinata	BO1	SRR4038982			
BRUCELLA_DEMO001	melitensis	16M			AE008917	
BRUCELLA_DEMO013	melitensis	63/9			CP007789	
BRUCELLA_DEMO014	melitensis	ATCC 23457			CP001488	
BRUCELLA_DEMO005	microti	CCM 4915			CP001578	
BRUCELLA_DEMO008	neotomae	5K33			JMSC01000001	
BRUCELLA_DEMO003	ovis	ATCC 25840			CP000709	
BRUCELLA_DEMO017	pinnipedialis	B2-94	SRR4038996			
BRUCELLA_DEMO002	suis	1330			CP002997	
BRUCELLA_DEMO011	vulpis	F60			LN997863	

- Input data
  - Experiment 'wgMLST'
    - Aspect
      TRUNK
  - Data type
    character data
- Network creation
  - Identification policy
    All taxa with zero inter-taxon distance have been identified.

Figure 12: The minimum spanning tree based on the TRUNK schema.

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***>.

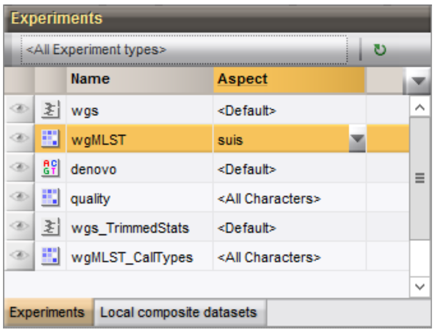
21. Highlight the *Comparisons* panel in the *Main* window and select ***Edit*** > ***Create new object...*** (✚) to create a new comparison for the selected entries.

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 13).

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



Name	Aspect
wgs	<Default>
wgMLST	suis
denovo	<Default>
quality	<All Characters>
wgs_TrimmedStats	<Default>
wgMLST_CallTypes	<All Characters>

Figure 13: The **suis** subschema.

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

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).

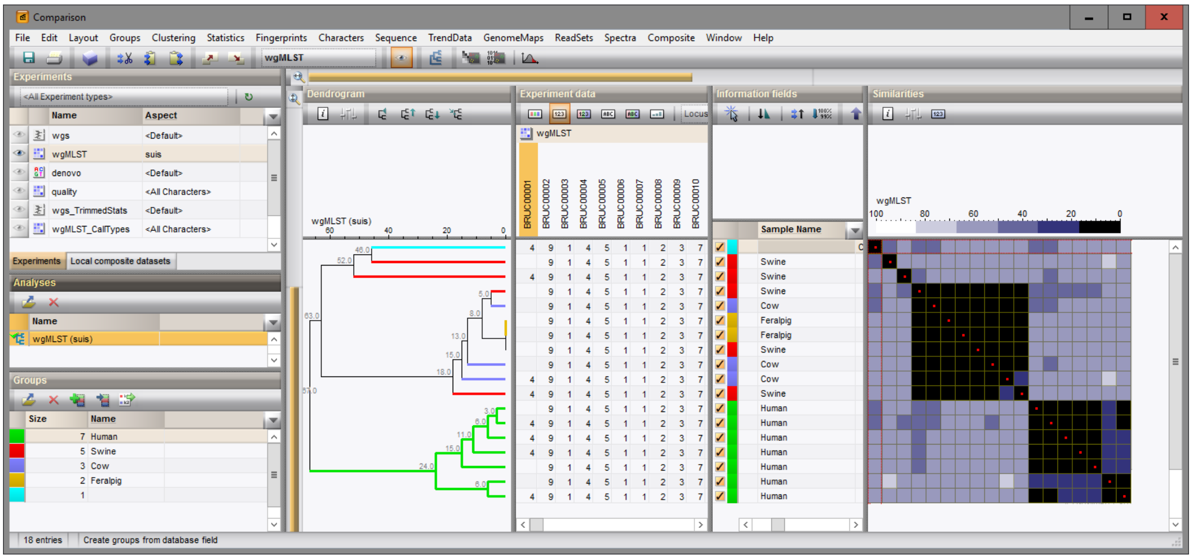

Comparison
File Edit Layout Groups Clustering Statistics Fingerprints Characters Sequence TrendData GenomeMaps ReadSets Spectra Composite Window Help
Experiments
Dendrogram
Experiment data
Information fields
Similarities
wgMLST (suis)
Sample Name
Swine
Cow
Feralpig
Human
Groups
7 Human
5 Swine
3 Cow
2 Feralpig
18 entries
Create groups from database field

Figure 14: 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 15).

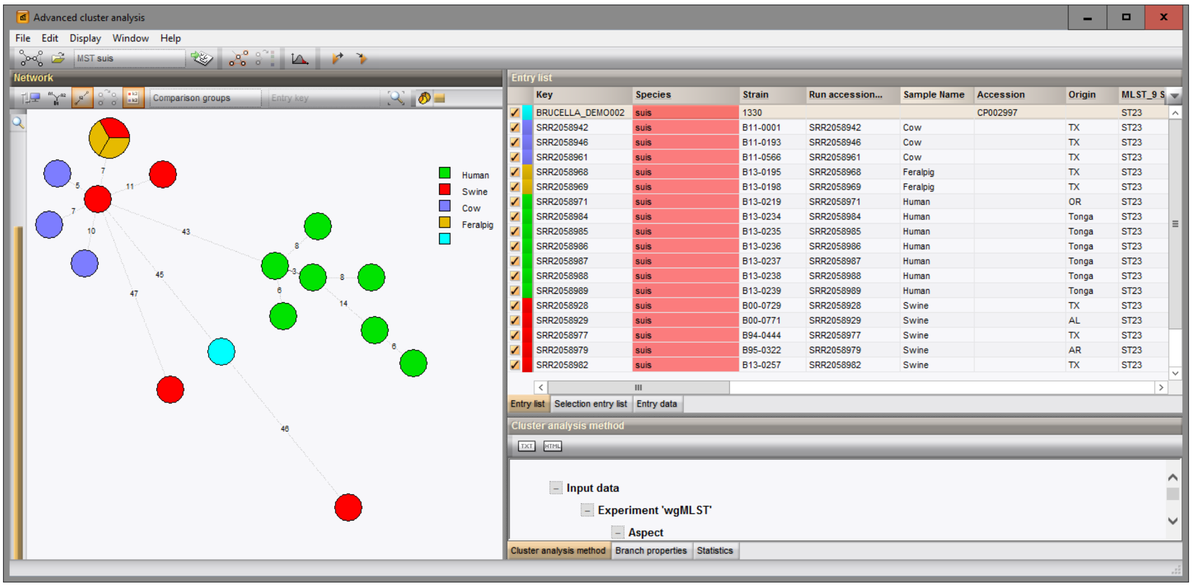
39. Close the *Advanced cluster analysis* window.

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.



		Key	Species	Strain	Run accession...	Sample Name	Accession	Origin	MLST_9 S
☑	🟦	BRUCELLA_DEMO002	suis	1330			CP002997		ST23
☑	🟪	SRR2058942	suis	B11-0001	SRR2058942	Cow		TX	ST23
☑	🟪	SRR2058946	suis	B11-0193	SRR2058946	Cow		TX	ST23
☑	🟪	SRR2058961	suis	B11-0566	SRR2058961	Cow		TX	ST23
☑	🟨	SRR2058968	suis	B13-0195	SRR2058968	Feralpig		TX	ST23
☑	🟨	SRR2058969	suis	B13-0198	SRR2058969	Feralpig		TX	ST23
☑	🟩	SRR2058971	suis	B13-0219	SRR2058971	Human		OR	ST23
☑	🟩	SRR2058984	suis	B13-0234	SRR2058984	Human		Tonga	ST23
☑	🟩	SRR2058985	suis	B13-0235	SRR2058985	Human		Tonga	ST23
☑	🟩	SRR2058986	suis	B13-0236	SRR2058986	Human		Tonga	ST23
☑	🟩	SRR2058987	suis	B13-0237	SRR2058987	Human		Tonga	ST23
☑	🟩	SRR2058988	suis	B13-0238	SRR2058988	Human		Tonga	ST23
☑	🟩	SRR2058989	suis	B13-0239	SRR2058989	Human		Tonga	ST23
☑	🟥	SRR2058928	suis	B00-0729	SRR2058928	Swine		TX	ST23
☑	🟥	SRR2058929	suis	B00-0771	SRR2058929	Swine		AL	ST23
☑	🟥	SRR2058977	suis	B94-0444	SRR2058977	Swine		TX	ST23
☑	🟥	SRR2058979	suis	B95-0322	SRR2058979	Swine		AR	ST23
☑	🟥	SRR2058982	suis	B13-0257	SRR2058982	Swine		TX	ST23

- Input data
  - Experiment 'wgMLST'
    - Aspect

Figure 15: The minimum spanning tree based on the **suis** subschema.

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 9).

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 16).

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.

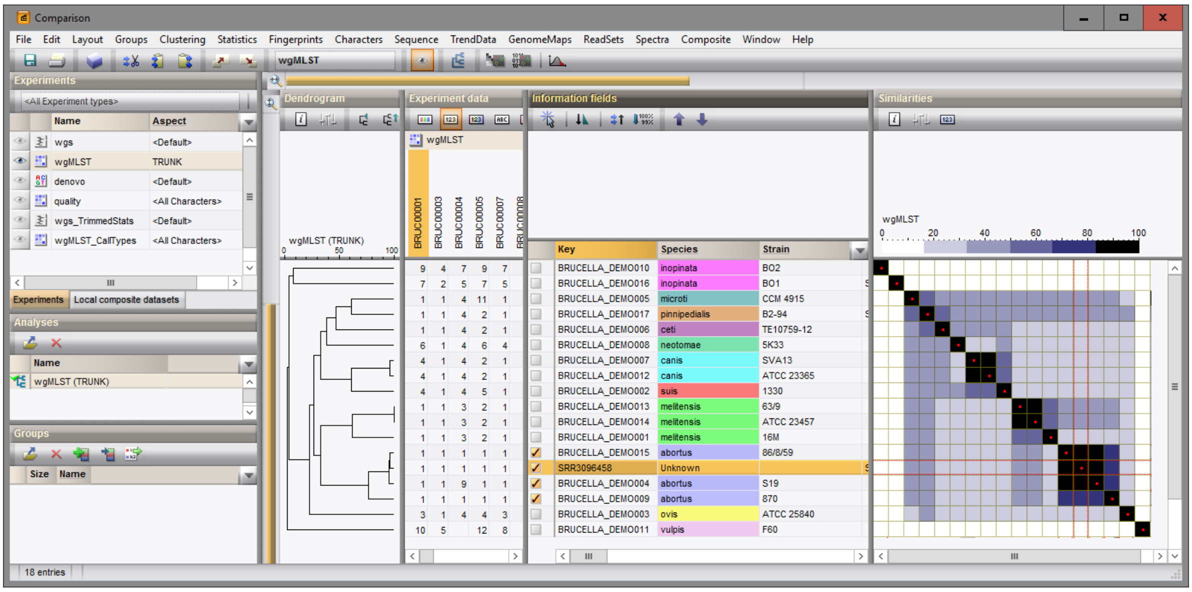
50. Select the **wgMLST** character experiment in the *Experiments* panel of the *Comparison* window.

51. Make sure the **abortus** aspect is selected for the **wgMLST** experiment (see Figure 17).

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***>.



Comparison
File Edit Layout Groups Clustering Statistics Fingerprints Characters Sequence TrendData GenomeMaps ReadSets Spectra Composite Window Help
wgMLST
Experiments
Dendrogram
Experiment data
Information fields
Similarities
wgMLST (TRUNK)
wgMLST
Analyses
Groups
18 entries

<All Experiment types>

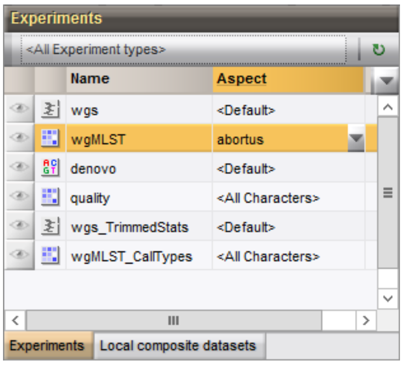
Name	Aspect
wgs	<Default>
wgMLST	TRUNK
denovo	<Default>
quality	<All Characters>
wgs_TrimmedStats	<Default>
wgMLST_CallTypes	<All Characters>

Experiments | Local composite datasets

Name
wgMLST (TRUNK)

BRUC00001	BRUC00003	BRUC00004	BRUC00005	BRUC00007		Key	Species	Strain
9	4	7	9	7	☐	BRUCELLA_DEMO010	inopinata	BO2
7	2	5	7	5	☐	BRUCELLA_DEMO016	inopinata	BO1
1	1	4	11	1	☐	BRUCELLA_DEMO005	microti	CCM 4915
1	1	4	2	1	☐	BRUCELLA_DEMO017	pinnipedialis	B2-94
1	1	4	2	1	☐	BRUCELLA_DEMO006	ceti	TE10759-12
6	1	4	6	4	☐	BRUCELLA_DEMO008	neotomae	5K33
4	1	4	2	1	☐	BRUCELLA_DEMO007	canis	SVA13
4	1	4	2	1	☐	BRUCELLA_DEMO012	canis	ATCC 23365
4	1	4	5	1	☐	BRUCELLA_DEMO002	suis	1330
1	1	3	2	1	☐	BRUCELLA_DEMO013	melitensis	63/9
1	1	3	2	1	☐	BRUCELLA_DEMO014	melitensis	ATCC 23457
1	1	3	2	1	☐	BRUCELLA_DEMO001	melitensis	16M
1	1	1	1	1	☑	BRUCELLA_DEMO015	abortus	86/8/59
1	1	1	1	1	☑	SRR3096458	Unknown	
1	1	9	1	1	☑	BRUCELLA_DEMO004	abortus	S19
1	1	1	1	1	☑	BRUCELLA_DEMO009	abortus	870
3	1	4	4	3	☐	BRUCELLA_DEMO003	ovis	ATCC 25840
10	5		12	8	☐	BRUCELLA_DEMO011	vulpis	F60

Figure 16: The *Brucella abortus* cluster.



<All Experiment types>

Name	Aspect
wgs	<Default>
wgMLST	abortus
denovo	<Default>
quality	<All Characters>
wgs_TrimmedStats	<Default>
wgMLST_CallTypes	<All Characters>

Experiments | Local composite datasets

Figure 17: The **abortus** aspect.

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).

55. Close the *Comparison* window.