BioNumerics Tutorial:

wgMLST typing in the Staphylococcus aureus 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 of *Staphylococcus aureus* from three studies, as they were published on NCBI's sequence read set archive.

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

2 Preparing the database

The **WGS demo database** for *Staphylococcus aureus* 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 Staphylococcus aureus** from the list and select ***Database*** > ***Download*** ().
3. Confirm the installation of the database and press <***Yes***> after successful installation of the database.
4. Close the *Tutorial databases* window with ***File*** > ***Exit***.

The **WGS demo database for Staphylococcus aureus** appears in the *BioNumerics Startup* window.

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

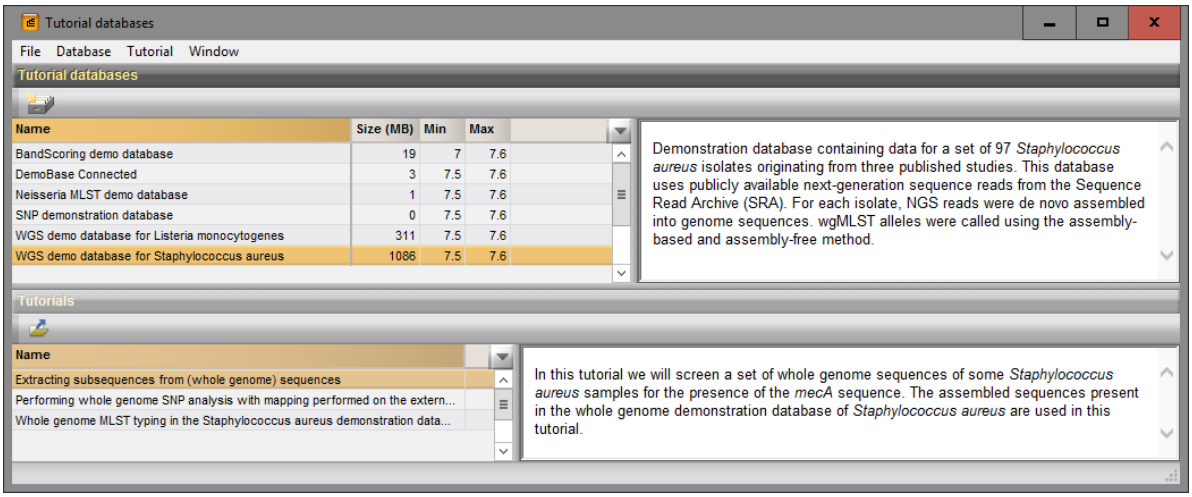

Figure 1: The *Tutorial databases* window, used to download the demonstration database.

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

A BioNumerics back-up file of the whole genome demo database for Staphylococcus aureus is also available on our website. This backup can be restored to a functional database in BioNumerics.

6. Download the file `wgMLST_SAUR.bnbk` file from `http://www.applied-maths.com/download/sample-data`, under 'WGS demo database for Staphylococcus aureus'.

In contrast to other browsers, some versions of Internet Explorer rename the `wgMLST_SAUR.bnbk` database backup file into `wgMLST_SAUR.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. From the menu that appears, select ***Restore database...***.

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

9. Specify a new name for this demonstration database, e.g. "Whole genome Staphylococcus aureus demobase".

10. Click <***OK***> to start restoring the database from the backup file (see Figure 2).

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

3 About the demonstration database

The demobase contains links to sequence read set data on NCBI's sequence read archive (SRA) for 97 publicly available sequencing runs of three *Staphylococcus aureus* whole genome sequencing studies ([1] [2] [3]) (see Figure 3). Sequence read set experiment type **wgs** contains the link to the sequence read set data on NCBI (SRA) with some raw data statistics.

The full wgMLST analysis (de novo assembly, assembly-based calls and assembly-free calls) was performed

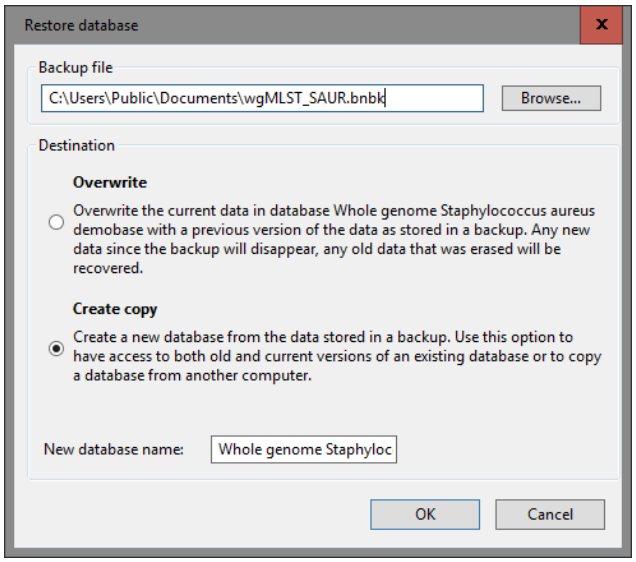

Figure 2: Restoring the whole genome demonstration database from the BioNumerics backup file `wg_SAUR.bnbk`.

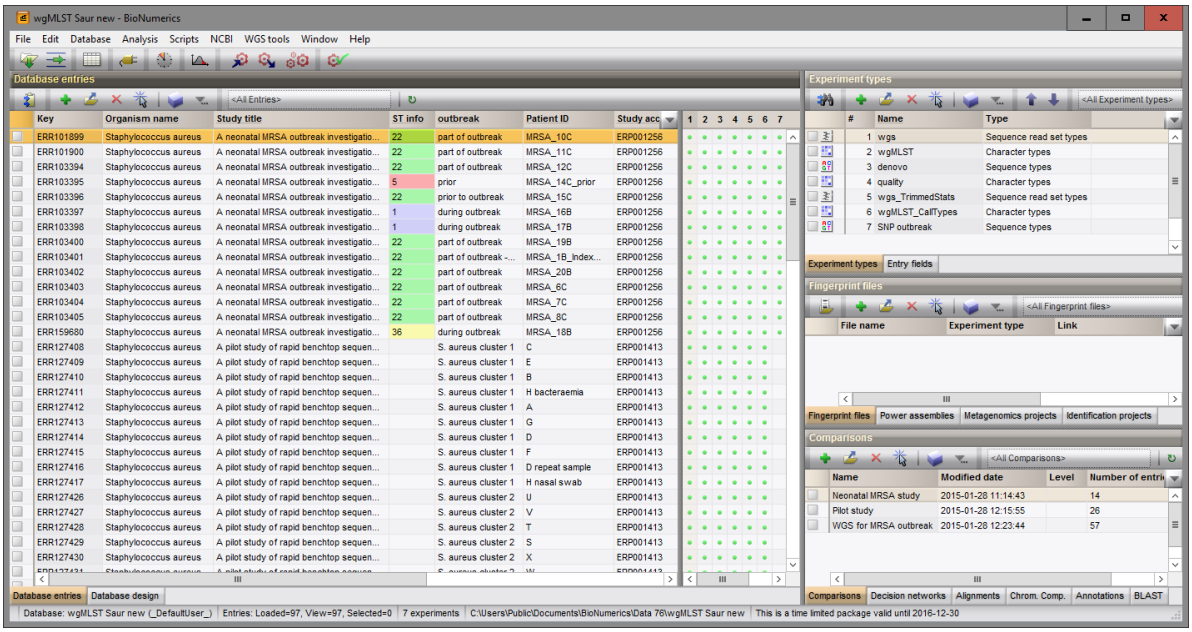

Figure 3: The *Staphylococcus aureus* demonstration database: the *Main* window.

on this set of samples using default settings and the *S. aureus* wgMLST scheme on the Applied Maths Cloud Calculation Engine.

1. Select ***WGS tools*** > ***Settings...***, click on the *wgMLST tab* (see Figure 4) and press the <***Auto submission criteria***> button (see Figure 5).

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.

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

Five experiment types linked to wgMLST are present in the database for each of the entries and are displayed in the *Experiment types* panel:

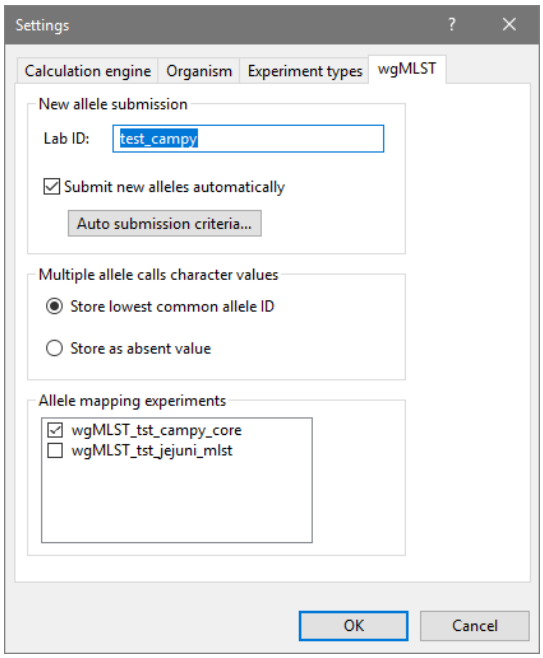

Figure 4: The *wgMLST tab* of the *Calculation engine settings* dialog box.

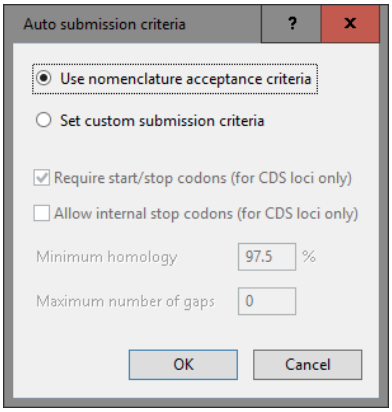

Figure 5: The *Auto submission criteria* dialog box.

- Character experiment type **wgMLST** contains the allele calls for detected loci in each sample, where the consensus from assembly-based and assembly-free calling resulted in a single allele ID.
- Sequence experiment type **denovo** contains the results from the de novo assembly algorithm, i.e. concatenated de novo contig sequences.
- Character experiment type **quality** contains quality statistics for the raw data, the de novo assembly and the different allele identification algorithms.
- Sequence read set experiment type **wgs_TrimmedStats**: contains some data statistics about the reads retained after trimming.
- Character experiment type **wgMLST_CallTypes**: contains details on the call types.

A reference mapping has been calculated for all entries from the Neonatal MRSA study and the resulting sequences are stored in the **SNP outbreak** sequence type. These sequences are used in the tutorial "Performing whole genome SNP analysis on Staphylococcus aureus genomes" to illustrate the wgSNP functionality present in BioNumerics.

Additional information (in entry info fields Organism name, Instrument, Study accession, etc.) was collected from the corresponding publications and added to the demonstration database. Additionally, a number of comparisons were created that include all the samples together or grouped per study.

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:

3. Click on the green colored dot for one of the entries in the first column in the *Experiment presence* panel. Column 1 corresponds to the first experiment type listed in the *Experiment types* panel, which is **wgs** in the default configuration.

In the *Sequence read set experiment* window, the link to the sequence read set data on NCBI (SRA) with a summary of the characteristics of the sequence read set is displayed: *Read set size*, *Sequence length statistics*, *Quality statistics*, *Base statistics* (see Figure 6).

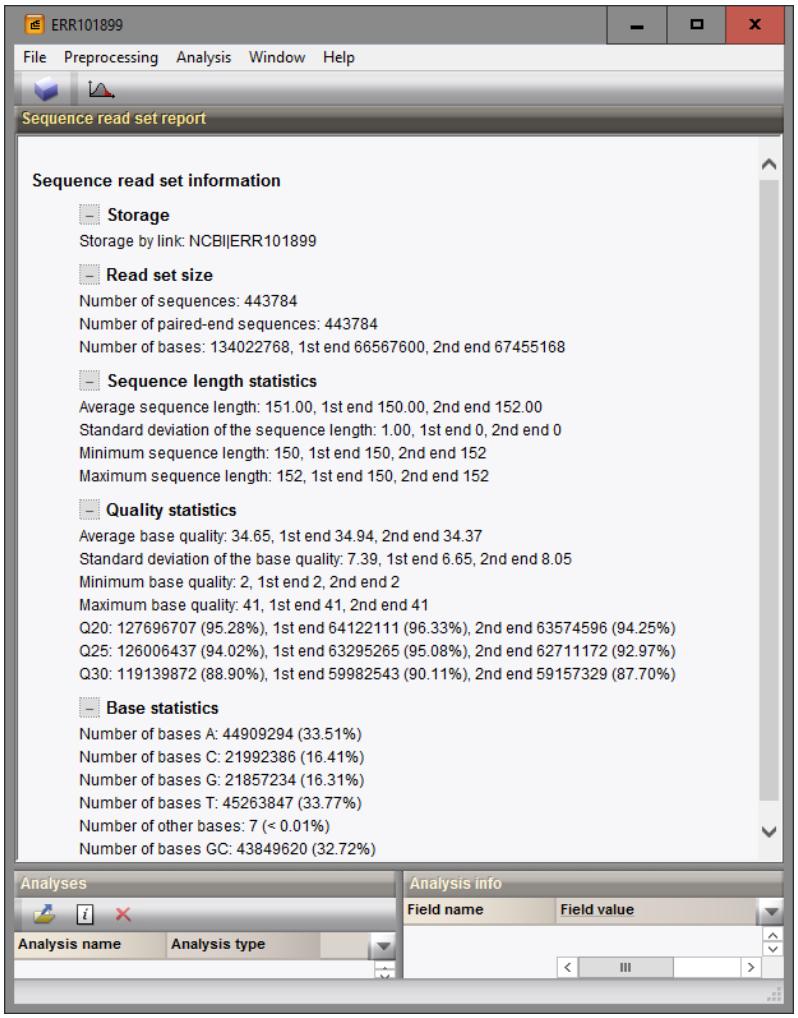

Figure 6: The sequence read set experiment card for an entry.

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

5. 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 resulted in a single allele ID (see Figure 7).

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

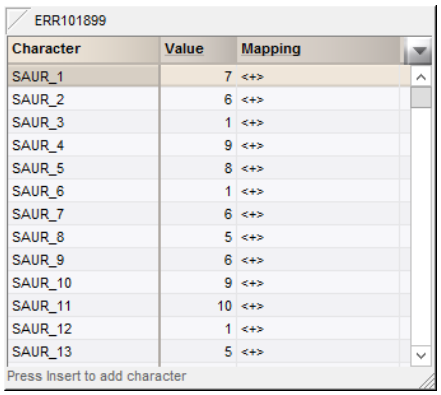

Figure 7: The character experiment card for an entry.

7. 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.

The *Sequence editor* window opens, containing the results from the de novo assembly algorithm, i.e. concatenated de novo contig sequences (see Figure 8).

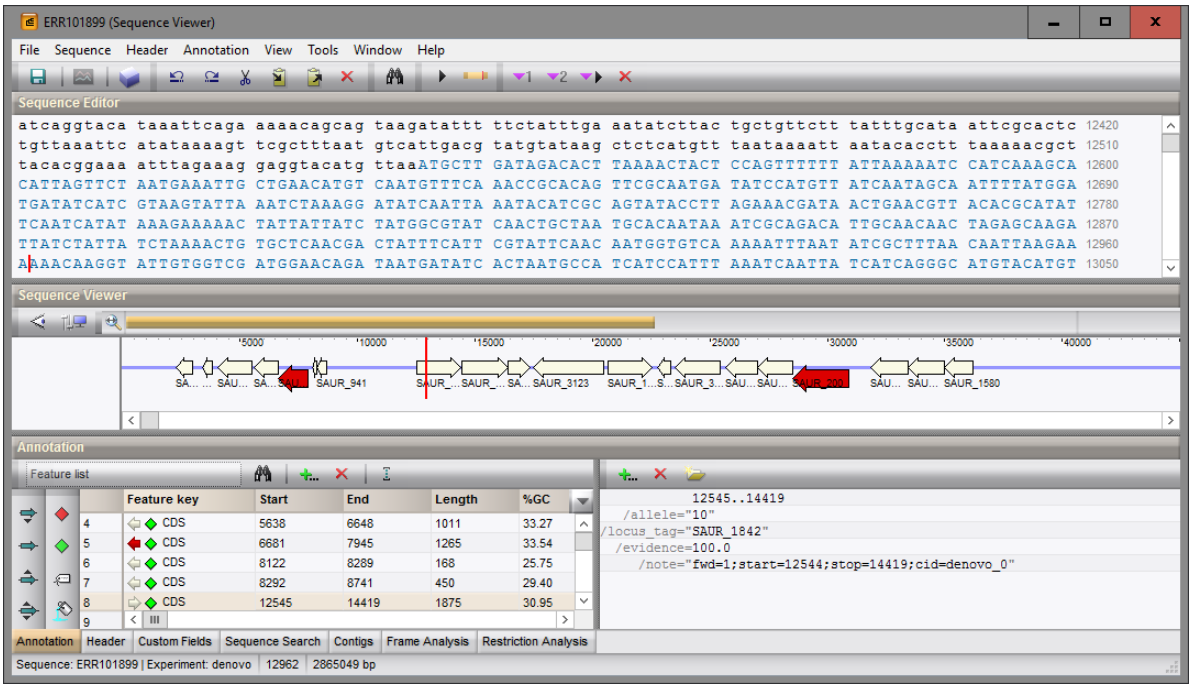

Figure 8: The *Sequence editor* window.

8. Close the *Sequence editor* window.

9. 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, the de novo assembly and the different allele identification algorithms (see Figure 9).

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

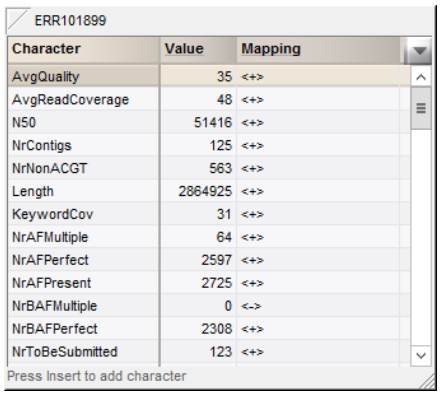

Figure 9: The character experiment card for an entry.

4 Subschemes

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

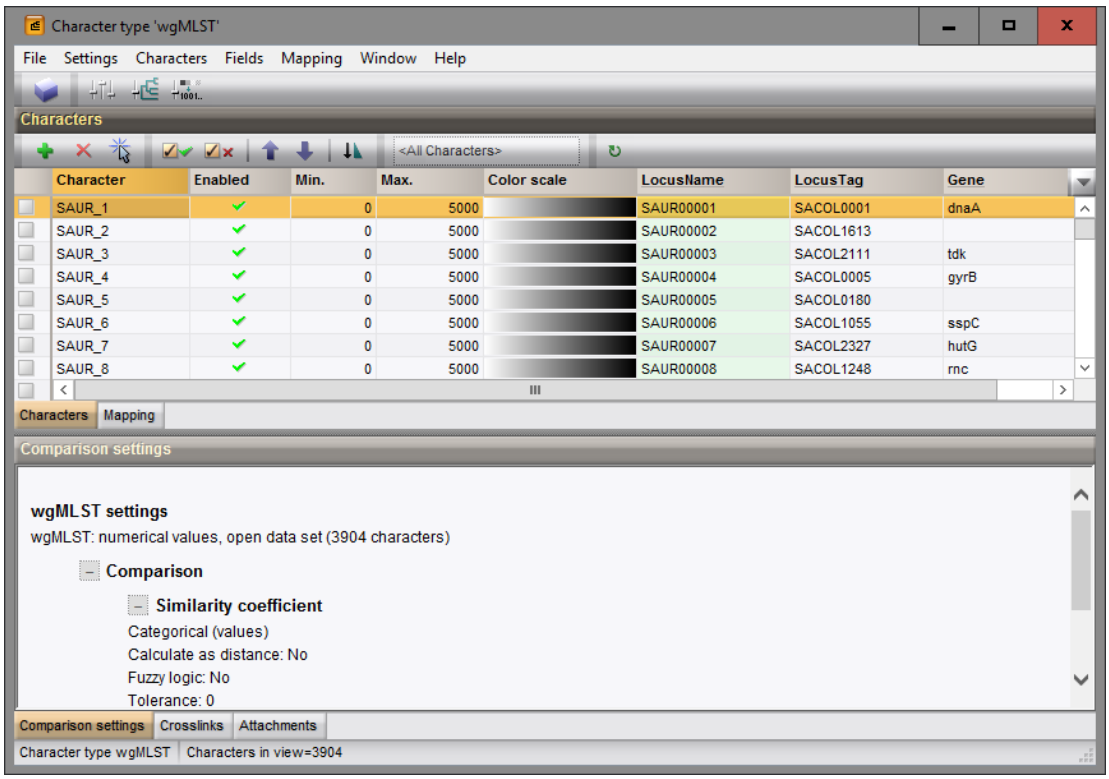

Figure 10: The *Character type* window.

Within a character experiment type, a character view can be defined that specifies a particular subset of characters.

2. Click on the drop-down bar in the toolbar (see Figure 11).

In this database, four views have been defined at the curator level and are synchronized upon installation: the default view **All loci**, the **MLST PubMLST** view for the traditional seven housekeeping loci, the **Core loci** view and the **wgMLST loci** view containing all loci except the ones present in the **MLST PubMLST** view.

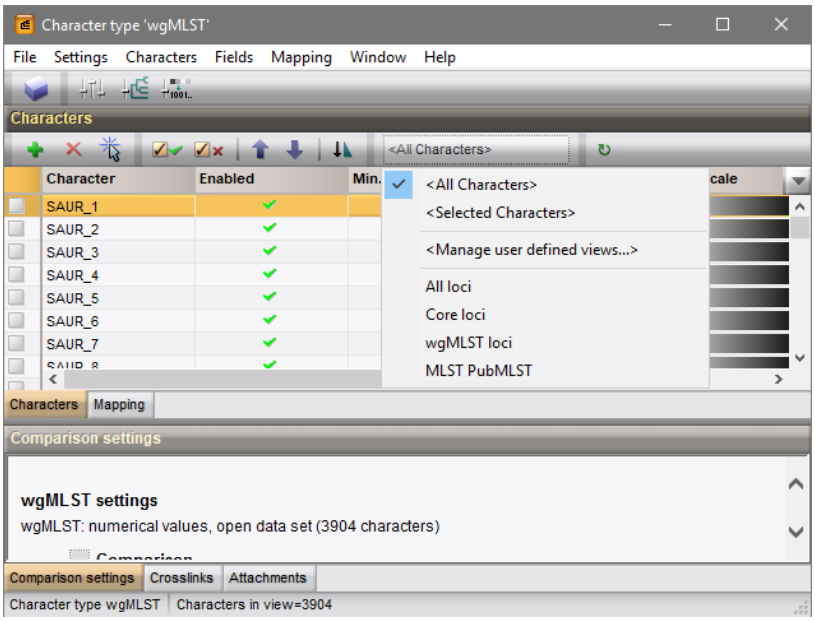

Figure 11: Views defined at the curator side.

3. Select the **MLST PubMLST** view from the list.

After selecting a character view, the window is updated (see Figure 12), and the number of characters in view is displayed in the status bar at the bottom of the window.

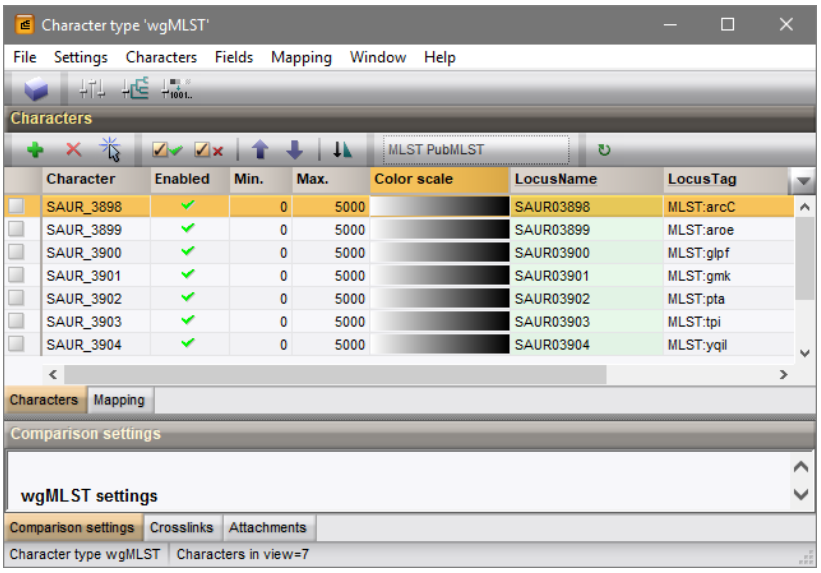

Figure 12: MLST loci from PubMLST.

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. Creating a character view can be done in two ways:

- The first method is based on a character *selection*.
- The second method is based on a *dynamic query* using the character information fields.

5. Select a few characters by selecting the characters directly in the *Character type* window (**Ctrl+click** or **Shift+click**).

The selection is synchronized with the database: any selection of characters made in the *Character type* window is reflected in other windows, e.g. the *Comparison* window, and vice versa.

6. Click on the drop-down bar in the toolbar and choose ***Manage user defined views***, alternatively select ***Characters*** > ***Character Views*** > ***Manage user defined views...*** (<All Characters>).

7. Press <***Add...***>, specify a name, e.g. **MySubsetExample**, make sure **Subset based** is selected, and press <***OK***> and <***Exit***>.

The new view is added to the database and is automatically selected in the *Character type* window. The new view is available for use e.g. in the *Character type* window, *wgMLST quality assessment* window or *Comparison* window.

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

As a second example we will create a query-based view of all loci encoding a ribosomal protein. Because all those loci have a gene name starting with ”rpl” (ribosomal proteins of the large subunit) or ”rps” (ribosomal proteins of the small subunit), this subset can be easily defined with a query-based view.

9. Click on the drop-down bar in the toolbar and choose ***Manage user defined views***, alternatively select ***Characters*** > ***Character Views*** > ***Manage user defined views...*** (<All Characters>).

10. Select <***Add...***>, specify a name, e.g. “ribosomal proteins”, make sure ***Query based*** is selected and click <***OK***>.

11. Select the ’Gene’ field, change the ***Equals*** condition to ***Contains*** and type “rpl” in the white box.

12. Press <***Add new***> in the ***Statements*** panel and edit it to ’Gene’ ***Contains*** “rps”.

13. Press <***Remove all unused***>.

14. Finally, select both remaining rules (use **Ctrl+click**) and press <***OR***> in the ***Group by*** panel.

The query should now look like in Figure 13.

15. Press <***OK***> to validate the query and <***Yes***> to confirm and press <***Exit***>.

The new query-based view is created with the 46 characters that fulfill the specified criteria (see Figure 14). The new view is available for use e.g. in the *Character type* window, *wgMLST quality assessment* window or *Comparison* window.

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

17. Close the *Character type* window.

5 Obtaining MLST profiles and sequence types

Using the *WGS tools plugin*, MLST profiles with public allele numbers can be obtained, i.e. using the same allele numbering as PubMLST. Additionally, the plugin allows the retrieval of public sequence types.

First, we need to activate the corresponding allele mapping experiment in the wgMLST settings:

1. Select ***WGS tools*** > ***Settings...*** to open the *Calculation engine settings* dialog box.

2. Click on the *wgMLST tab* to bring the wgMLST settings into focus.

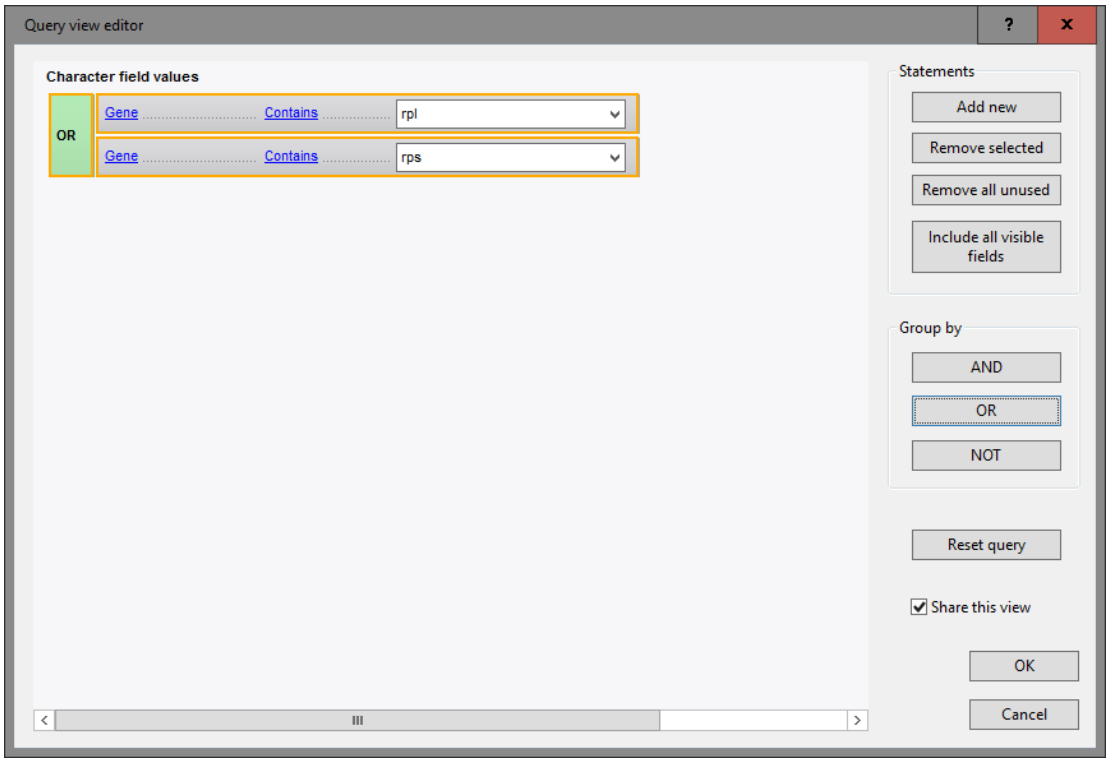

Figure 13: Query based view.

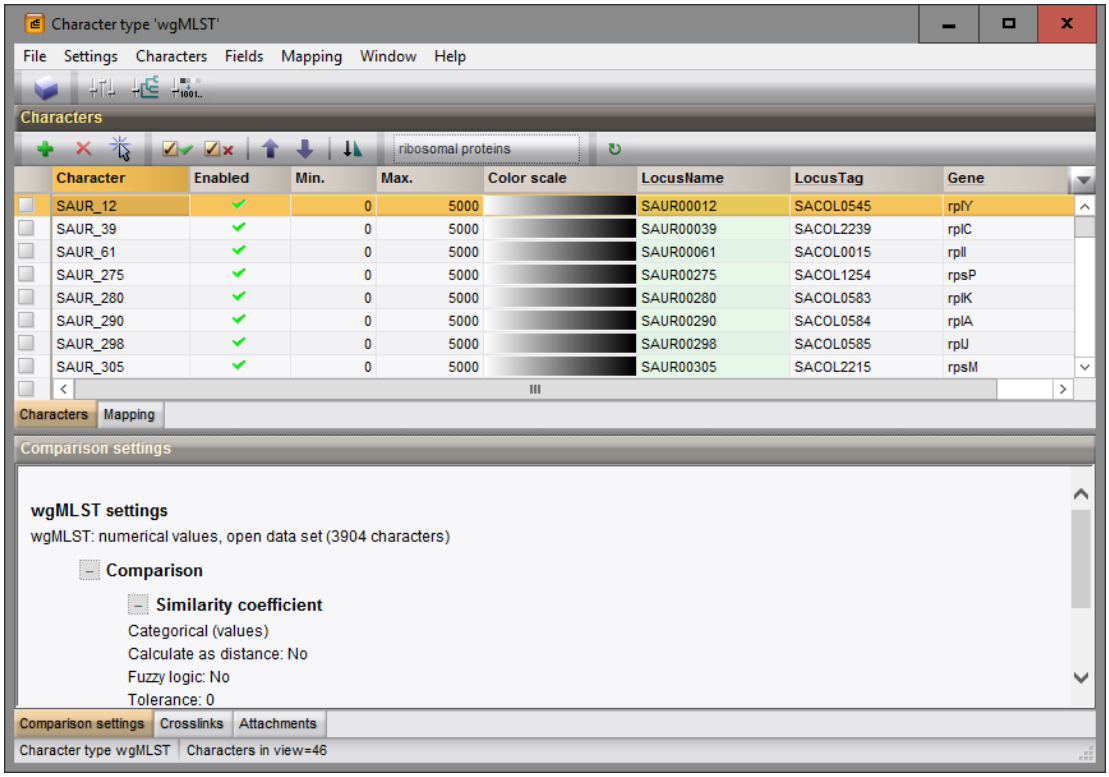

Figure 14: New query based view.

3. Under ***Allele mapping experiments***, check ***wgMLST_MLST PubMLST*** and press <***OK***>.

A character experiment type called **wgMLST_MLST PubMLST** is created in the database in case it did

not exist yet. Now, MLST profiles with exactly the same allele IDs as used on PubMLST can be obtained for all entries with a **wgMLST** experiment:

4. In the *Experiment types* panel, highlight the **wgMLST** experiment type and select ***Database*** > ***Entries*** > ***Select entries with experiment*** to make the entry selection.

5. Select ***WGS tools*** > ***Get alleles mapping***.

The allele numbers from the **wgMLST** experiments will be submitted to the Calculation Engine, where they are translated into public nomenclature. The public allele numbers are then retrieved and stored in the **wgMLST_MLST PubMLST** experiments. Optionally, this can be verified in the *Comparison* window:

6. Highlight the *Comparisons* panel and select ***Edit*** > ***Create new object...*** ( ) to open a comparison with the selected entries.

7. In the *Experiments* panel, click on the icon next to **wgMLST_MLST PubMLST** to visualize the MLST profiles in the *Experiment data* panel.

8. Close the *Comparison* window.

Next, sequence types can be assigned for the selected entries, based on the **MLST PubMLST** subscheme.

9. In the *Main* window, select ***WGS tools*** > ***Assign wgMLST sequence types...***.

This opens the *Assign sequence types* dialog box, where available typing schemes can be checked to be included in the assignment of the sequence types (see Figure 15).

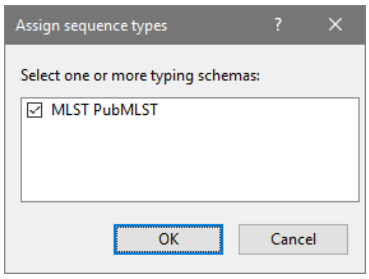

Figure 15: The *Assign sequence types* dialog box, with a single typing scheme listed.

10. Leave the subscheme **MLST PubMLST** checked and press <***OK***> to assign a sequence typing based on the 7 loci used for traditional MLST analysis.

Per entry and typing scheme, a list of allele identifications is sent to the allele database and sequence type information is returned. The sequence types are then saved to a dedicated entry information field.

In our example database, a sequence type is added in the field 'MLST PubMLST ST' for the selected entries.

In case an entry has an incomplete profile for the **MLST PubMLST** subscheme, no sequence type can be assigned and an error message will be generated for that entry.

6 Import of sample-specific allele sequences into the database

Once the wgMLST allele results have been imported in the database, it is possible to import the actual allele sequences for a specific wgMLST locus or a combination of loci, as defined in a subscheme, using ***WGS tools*** > ***Store wgMLST locus sequences...***.

As an example, we will describe how to retrieve the allele sequences for the seven MLST loci into the database, using sequence type names that can be recognized by the *MLST online plugin*. First, a character

info field should be created and the exact locus names as defined in the MLST scheme should be entered for those seven loci.

1. Open the **wgMLST** *Character type* window by double-clicking the character experiment type in the *Experiment types* panel (top right of *Main* window).

2. In the character views drop down menu, select ***MLST PubMLST***.

3. Fill in the names of the seven MLST loci as they are defined in the *S. aureus* MLST scheme on `http://saureus.mlst.net/`, in the new 'Gene' field: "arcc", "aroe", "glpf", "gmk", "pta", "tpi", and "yqil" (see Figure 16). A field becomes editable by clicking it after it was selected (click twice slowly).

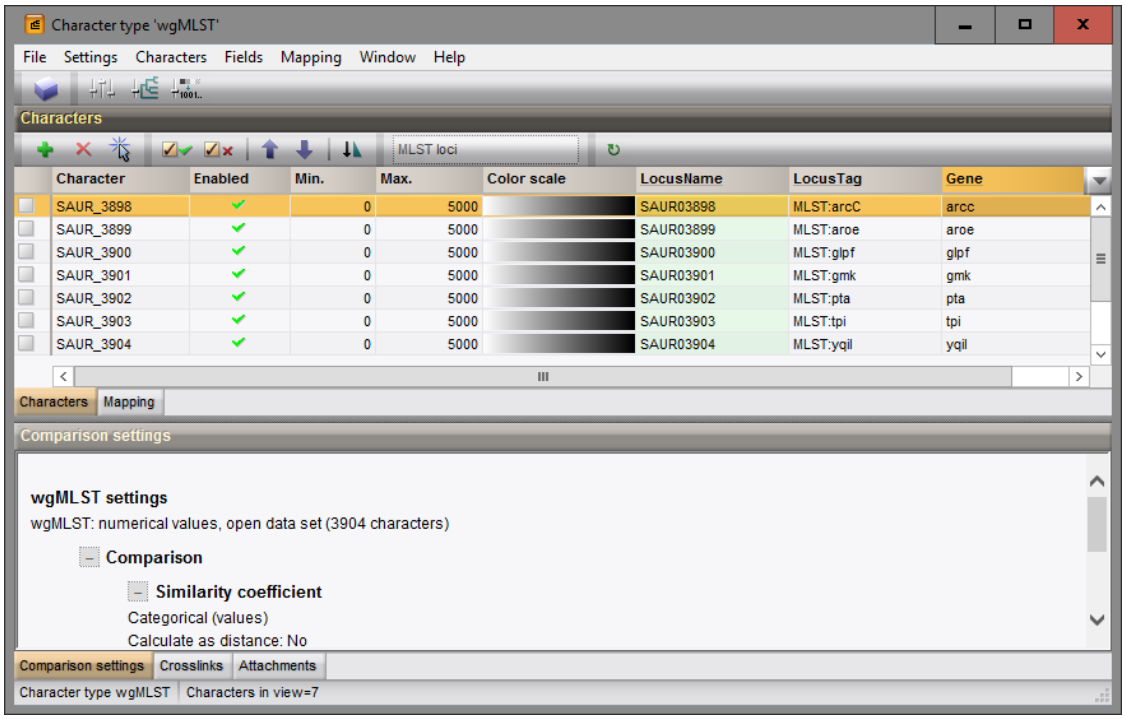

Figure 16: The *Character type* window for **wgMLST**, with locus names for the 7 MLST loci, as known on PubMLST.net, filled in in the 'Gene' character information field.

4. Close the *Character type* window.

Now the allele sequences can be imported into sequence type experiments that have the correct name for analysis by the *MLST online plugin*.

5. Make sure the *Database entries* panel is the active panel and select ***Edit*** > ***Select all*** (**Ctrl+A**) to select all entries at once.

6. Select ***WGS tools*** > ***Store wgMLST locus sequences...*** (see Figure 17). Specify "MLST PubMLST" as the ***Subschema*** and select "Gene" for the ***Sequence experiment type***.

7. Click <***OK***> to start importing the allele sequences and <***Yes***> to confirm the creation of new experiment types.

The database now contains the allele sequences for the 7 MLST loci, stored in 7 sequence experiment types that can be accessed by the *MLST online plugin*. This can be illustrated as follows:

8. Install the *MLST online plugin*, via ***File*** > ***Install / remove plugins...*** ( ). Select ***MLST online***, press <***Activate***> and confirm with <***OK***>.

9. Choose ***Select organism from online list*** and select **Staphylococcus aureus** from the list. Leave all the other settings at default: press <***Next***> several times, then <***Finish***> and confirm with <***OK***>. Close the *Plugins* dialog box.

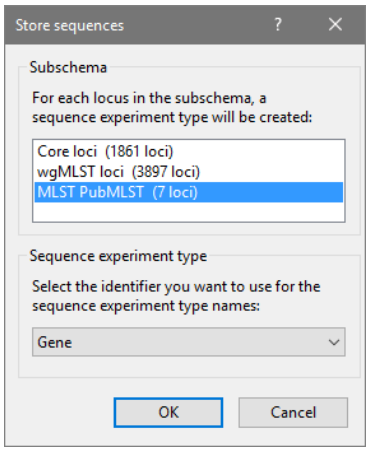

Figure 17: The *Store sequences* dialog box.

10. In the *Main* window, select all the entries via ***Edit*** > ***Select all*** (**Ctrl+A**) and choose ***MLST*** > ***Identify alleles and profiles***.

The character type **MLST** now contains the allele numbers for the 7 loci as they are known in the public PubMLST scheme, the public sequence types are written to the entry field **MLST ST** (see Figure 18). For two entries, one of the loci was not called, so no sequence was stored in the database and no sequence type could be assigned.

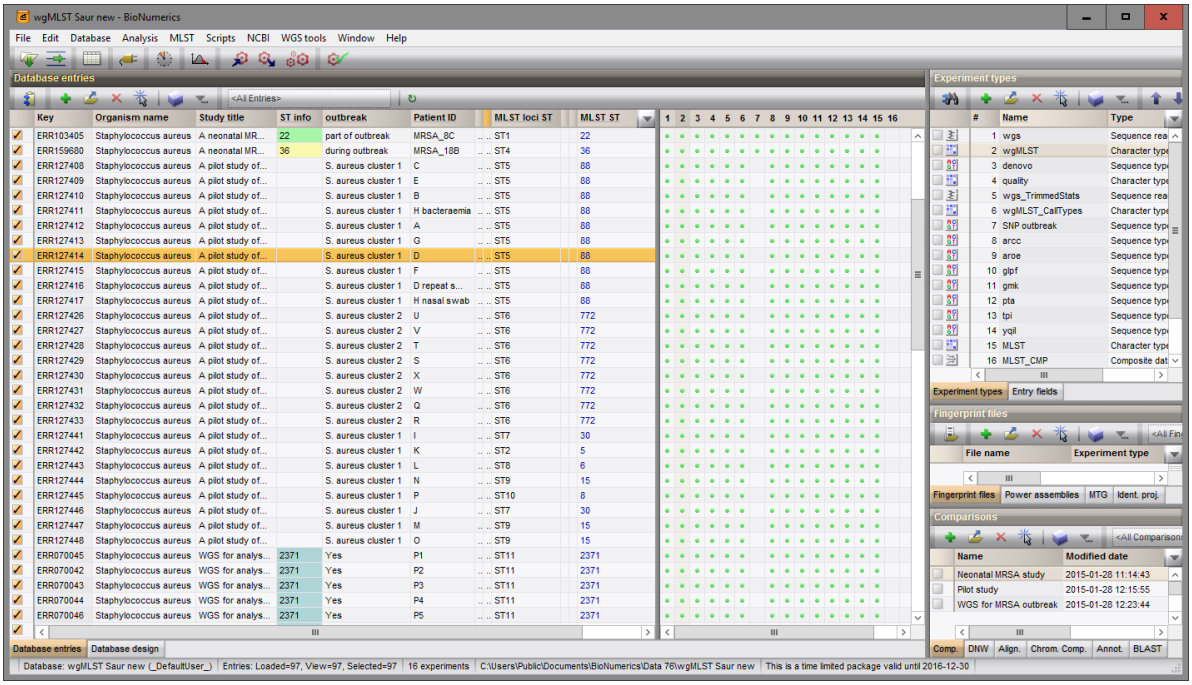

Figure 18: The *Main* window.

11. Click on the green colored dot for one of the entries in the **MLST** column in the *Experiment presence* panel.

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

Please consult the *MLST online plugin* manual for detailed instructions on how to proceed to submit the alleles to PubMLST and obtain public MLST sequence types.

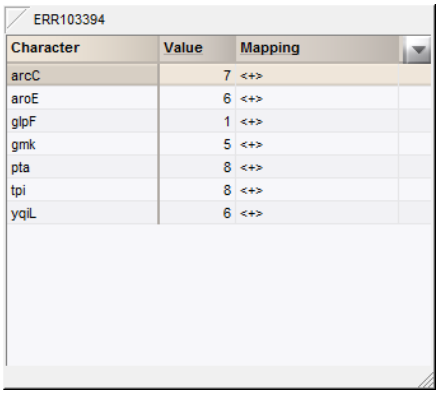

Figure 19: The MLST character experiment card.

7 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. We will detail here how a dendrogram and minimum spanning tree (MST) can be created from the *Comparison* window and the *Advanced cluster analysis* window, using data from [2].

7.1 Comparison window

In the WGS demonstration database, three comparisons are already created, corresponding to the three studies (see Figure 20).

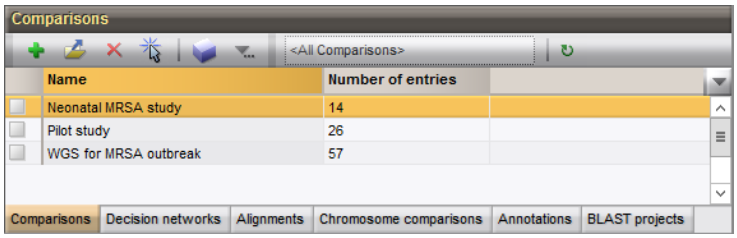

Figure 20: The *Comparisons* panel with the three comparisons.

Creating a new comparison is easily achieved by selecting the entries you would like to include in the *Main* window and clicking on the icon in the *Comparisons* panel. Here we will work with the selection of entries present in the saved **WGS for MRSA outbreak** comparison:

1. Open comparison **WGS for MRSA outbreak** by double-clicking it in the *Comparisons* panel in the *Main* window.

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

A valuable addition in the analysis of wgMLST data is the use of character views, i.e. wgMLST subschemes consisting of a subset of loci for a specific research question. Default **All characters** are included in the analysis. Another character view can be selected from the drop-down list in the **Aspect** column (see Figure 21).

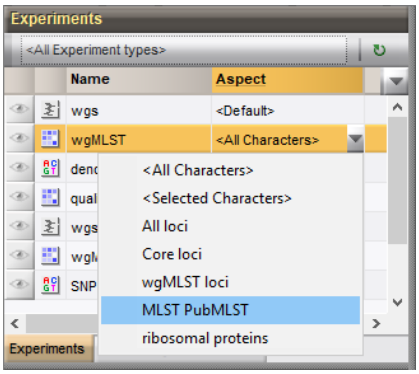

Figure 21: Character views in the **wgMLST** experiment type.

7.2 Similarity based clustering

The **WGS for MRSA outbreak** comparison contains saved cluster analyses, stored in the *Analyses* panel. The experiment and subscheme (between brackets) are indicated (e.g. **wgMLST (Core loci)**).

3. Switching between the analyses can be done by double-clicking them from the *Analyses* panel.

As an example we will perform a new cluster analysis, only based on the 7 traditional MLST loci.

4. Select the **MLST PubMLST** character view of the **wgMLST** character experiment in the *Experiments* panel.

5. In the *Experiments* panel click on the eye icon ( ) that proceeds **wgMLST** to display the values of the 7 MLST loci.

6. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***, select ***Categorical (values)***, make sure ***Calculate as distance*** is unchecked, press <***Next***>, choose ***UPGMA*** 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 (see Figure 22). The subscheme that was used is indicated between brackets: **wgMLST (MLST PubMLST)**.

From the dendrogram it is clear that all the samples with ST 2371 cluster closely together. All these samples were isolated as part of an outbreak, either from patients or from one of the health care workers in the same facility. We will now calculate a dendrogram based on the core loci (alternatively double-click on the saved analysis **wgMLST (Core loci)** in the *Analyses* panel):

7. Select the **Core loci** character view of the **wgMLST** character experiment in the *Experiments* panel.

8. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...*** to start a cluster analysis.

9. Select the ***Categorical (values)*** similarity coefficient, press <***Next***>, and select the ***UPGMA*** clustering method. Press <***Finish***> to start the calculation of the dendrogram.

The resulting dendrogram is displayed in the *Dendrogram* panel. It is clear that the core loci provide a much higher resolution over the MLST set.

To study the relationships in ST 2371 cluster more closely, we can create a new comparison that includes only those entries. We can select only the entries from within the comparison by emptying the current selection and then clicking on the node that contains all the ST 2371 entries while holding the **Ctrl**-key. Alternatively, we can make a new selection in the *Main* window.

10. In the *Main* window, clear the current selection with ***Database*** > ***Entries*** > ***Unselect all entries (all levels)*** (, **F4**), then use ***Edit*** > ***Find object in list...*** (, **Ctrl+Shift+F**) to open the *Find* dialog box.

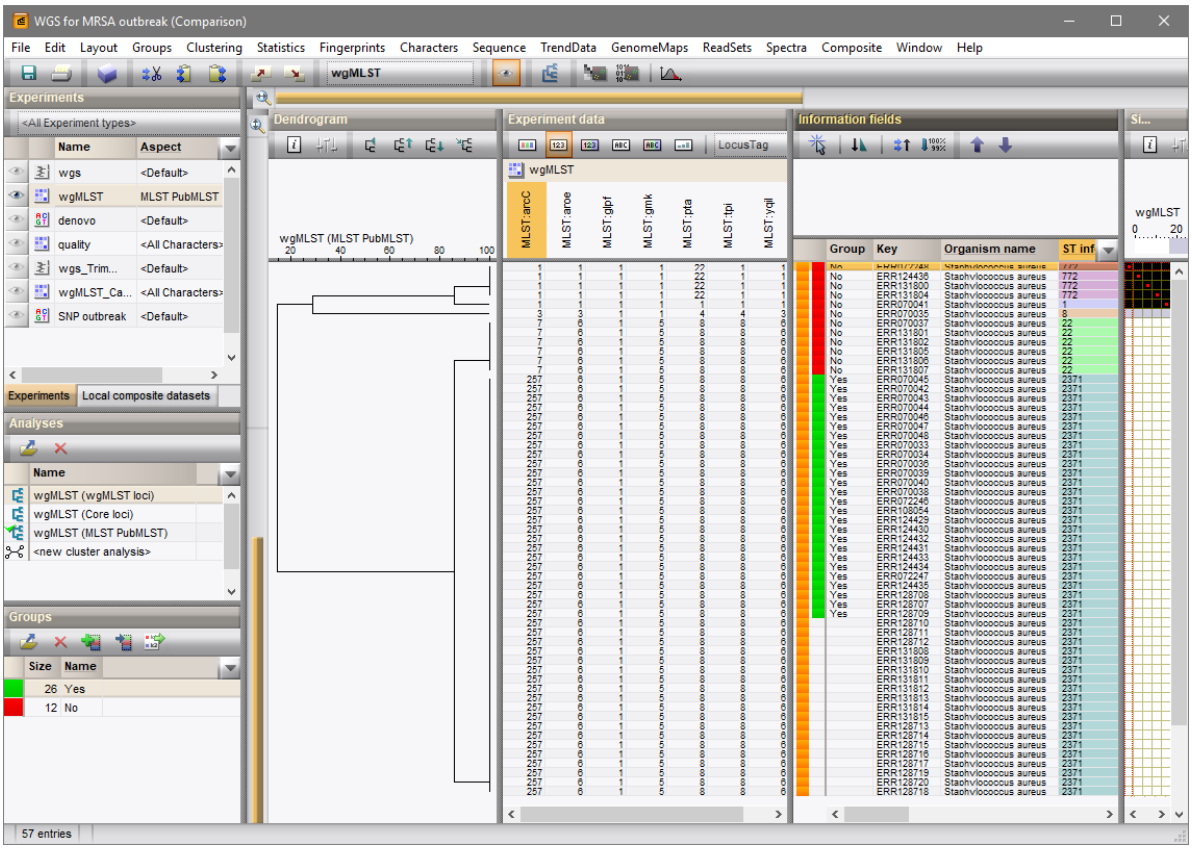

Figure 22: Dendrogram based on the MLST loci.

11. Type “2371” and press <***Select all***> to select the 45 entries.

12. Create a new comparison for the selected entries by clicking on the icon in the *Comparisons* panel.

We can add some information to the MST we are about to create, by specifying comparison groups. In the database, samples isolated from a patient have the label ”Yes” in the field ***Outbreak***, whereas the samples isolated from a health care worker do not carry a label:

13. Right-click on the **Outbreak** column header 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***>.

14. Select the **wgMLST loci** aspect for **wgMLST** in the *Experiments* panel.

15. In the *Experiments* panel click on the eye icon ( ) that proceeds **wgMLST** to display the values of the wgMLST loci.

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

A disadvantage of the ***Categorical (values)*** similarity coefficient is that the number of different loci cannot easily be deduced from the dendrogram or similarity matrix. The ***Categorical (differences)*** coefficient is more suitable for this purpose.

17. Select the ***Categorical (differences)*** coefficient from the list.

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.

18. In this example, choose a ***Scaling factor*** of **1**.

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

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

20. 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 (in this example: 1).

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

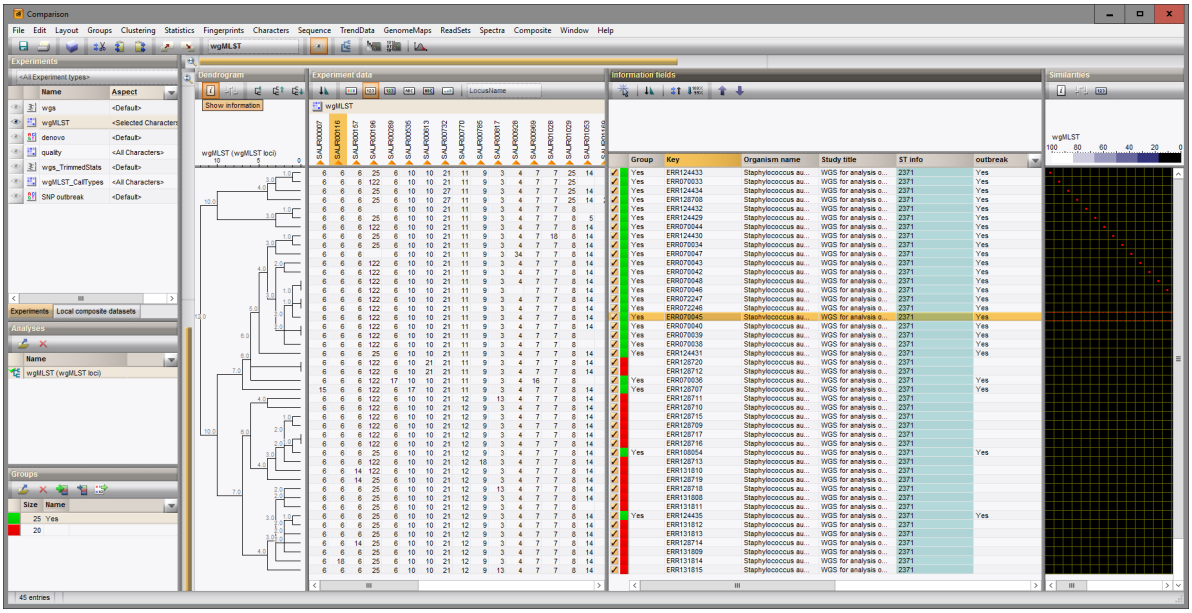

Figure 23: Complete linkage dendrogram.

22. Save the comparison with ***File*** > ***Save as...***. Specify a name (e.g. **ST 2371**).

7.3 Minimum spanning tree

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

23. Open the saved comparison **ST 2371** or create a new comparison containing all 45 entries in the database belonging to ST 2371.

24. Select the **wgMLST loci** character view of the **wgMLST** character experiment in the *Experiments* panel.

25. 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 with single and double locus

variance priority rules. We could use this template but as an example, we will demonstrate here how to create a new template:

26. Specify an analysis name (for example **wgMLST MST**), make sure ***wgMLST (wgMLST loci)*** is selected, select ***No template***, and press <***Next***>.

27. In the next step, leave the settings at default (***Character data***, ***Treat characters as categorical***, ***Merge taxa when distance is zero***), click <***Next***>.

28. For the ***Network creation algorithm***, choose ***Minimum Spanning Tree with hypothetical nodes*** and click <***Next***>.

29. Leave the algorithm details at their default values and press <***Next***>.

30. In the last page of the *Create network* wizard, check ***Prefer compact network*** and press <***Next***>.

A MST is now computed in the *Advanced cluster analysis* window.

31. To add more information to the MST, go to ***Display*** > ***Display settings***. In the *Node labels and sizes panel* of the *Display settings* dialog box, check ***Show node labels***. Choose "Patient ID" in the drop-down list and leave the other settings at their default values.

32. In the *Branch labels and sizes panel* of the *Display settings* dialog box, we can specify that we 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".

33. Click <***OK***> to close the *Display settings* dialog box.

The MST is now displayed with node and branch labels.

34. Zooming can be done with the zoom slider on the left side of the image, and the size of the nodes can be adjusted with the zoom slider at the top. By holding the **Ctrl**-key and dragging a node with the mouse, the node can be repositioned in any direction.

35. Export the image via ***File*** > ***Export image...*** and save in the format of your choice.

The resulting MST gives a high resolution map of the outbreak. The colors allow to distinguish easily between patient samples (P) and MRSA colonies isolated from a health care worker (HC).

The branch labels indicate how many allele differences were found between each linked set of entries. The conclusions from the whole genome SNP analysis performed in [2] are largely supported by the wgMLST analysis performed here.

By repeating the analysis steps using the character aspect "Core loci", it can be demonstrated that wgMLST results in a higher-resolution MST then core genome MLST.

8 Core and pan genome analysis

The pan-genome of a bacterial species consists of a core and an accessory gene pool. As the wgMLST locus set is defined as pan-genomics scheme over all available organism genome sequences, the analysis can be limited to the pan-genomic and/or core genomic loci for the selected sample set in the comparison.

For a selected set of samples, the core set of loci can be defined as follows:

1. Select all entries in the *Main* window and click on the  icon in the *Comparisons* panel.

2. In the *Experiments* panel of the *Comparison* window, highlight the **wgMLST** character experiment, make sure the "<All characters>" aspect is selected and select ***Statistics*** > ***Core locus analysis...***.

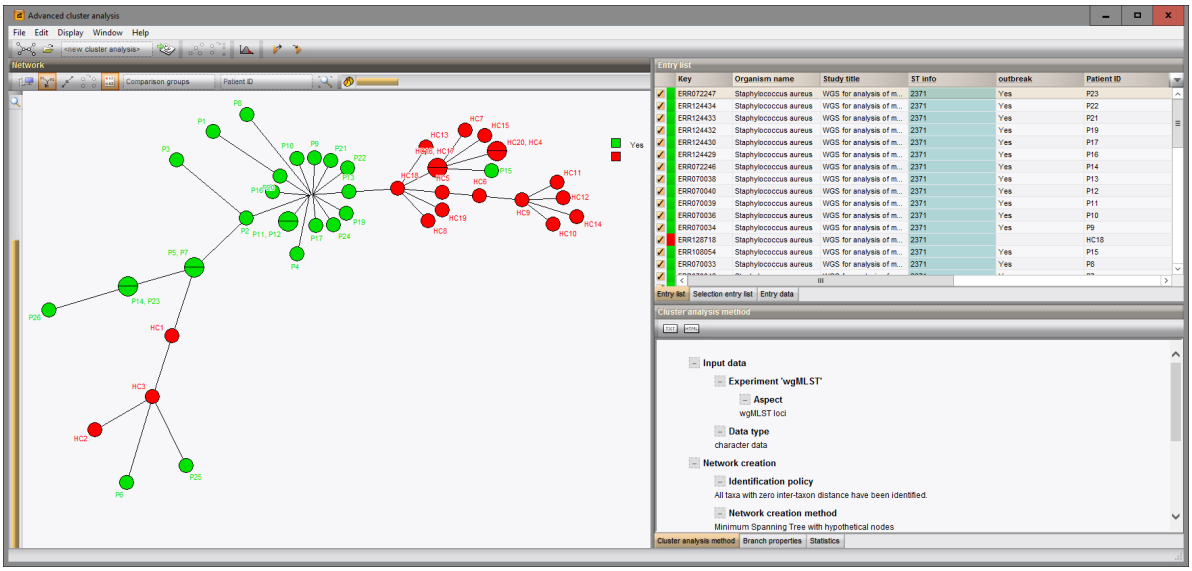

Figure 24: MST with hypothetical nodes of the entries with ST2371, after wgMLST analysis on the data described in [2].

This opens the *Core locus analysis* dialog box where the ***Number of repeats*** and ***Presence threshold*** can be defined.

The determination of the number of core loci is based on sub-sampling the entries in the comparison. As such, the ***Number of repeats*** can be defined, i.e. the number of subsamples taken from the comparison set.

The ***Presence threshold*** indicates the minimum presence (expressed in %) for a locus to be called within the core. Entering 90%, will imply that only loci present in 90% of the entry selection will be identified as core loci. For a very strict analysis, one can put the presence threshold at 100%, limiting the core to only those loci which are present in all the entries under evaluation i.e. present in the comparison.

3. When the analysis has finished, the results open in the *Charts and statistics* window.

4. To create a Core genome analysis plot as shown in Figure 25, highlight ***Average number of loci*** and select ***Plot*** > ***Add new plot from selected properties...*** (), choose ***Profile chart***.

5. Repeat Instruction 4 for data sources ***Minimum number of loci*** and ***Maximum number of loci***.

The result is shown in Figure 25.

6. The values used to create these curves can be viewed by making a selection of all data sources (click while holding the **Ctrl**-key) and selecting ***Dataset*** > ***View selected properties*** ().

For details on all the possibilities of the *Charts and statistics* window, please consult the BioNumerics reference manual.

7. The core loci are now also selected in the **wgMLST** character experiment, in the form of a subset-based character view (see 4).

8. Double-click on the **wgMLST** character experiment in the *Experiment types* panel of the *Main* window, create a selection based query, and specify a name that is different from the pre-defined **Core loci** sub-scheme, e.g. **Local core loci**.

The predefined Core loci subscheme consists of the 1861 loci found to be present in all of 39 reference sequences, as described in 4.

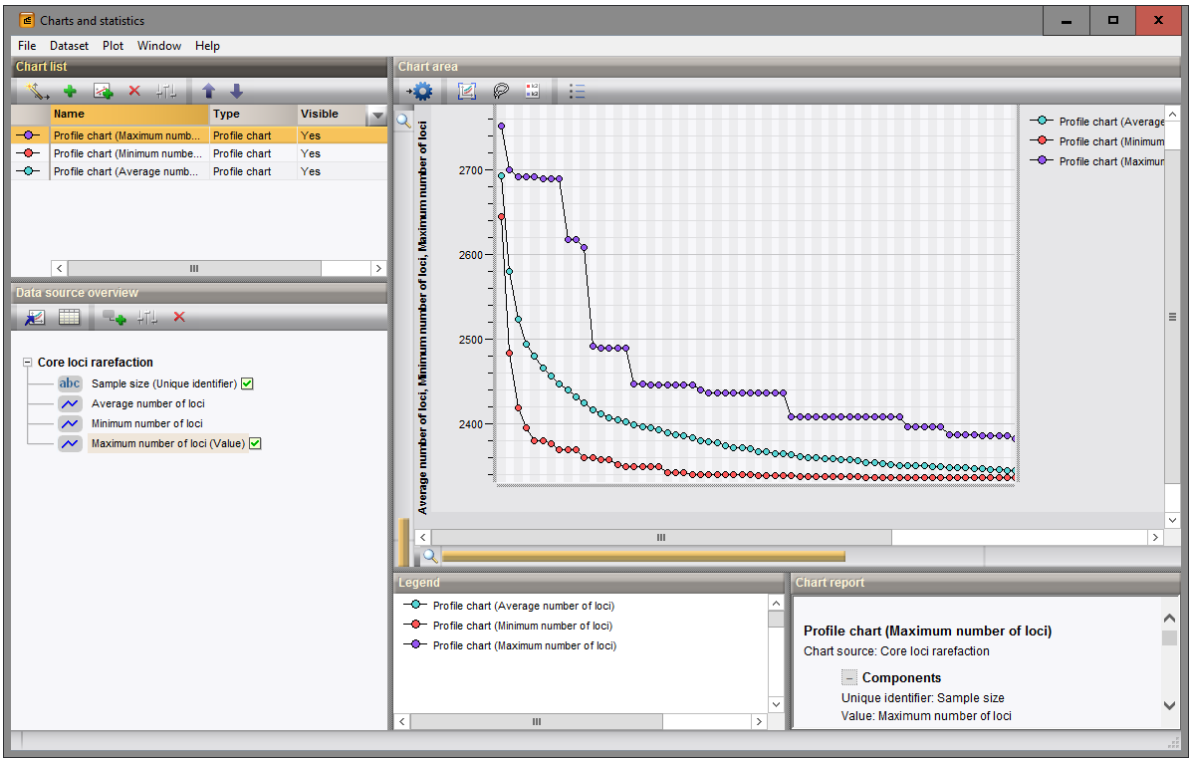

Figure 25: Core locus analysis for all samples in the wgMLST demonstration database (***Presence threshold*** 100%).

From the same *Comparison* window, also a pan locus analysis can be done.

9. In the *Comparison* window select ***Statistics*** > ***Pan locus analysis...***. As for the Core locus analysis, the ***Number of repeats*** and ***Presence threshold*** can be defined from the *Pan locus analysis* dialog box.

Similar to the determination of the number of core loci, the number of pan loci is also based on sub-sampling the entries in the comparison. As such, the ***Number of repeats*** can be defined, i.e. the number of subsamples taken from the comparison set.

The ***Presence threshold*** indicates the minimum presence (expressed in %) for a locus to be called within the pan loci. Entering 5%, will imply that only loci present in at least 5% of the selected entries will be identified as pan loci. For a very non-restrictive analysis, one can put the presence threshold at 0%, defining the pan loci as all the loci which are present in at least one of the entries.

10. When the analysis has finished, the results open in the *Charts and statistics* window.

11. To create a Pan genome analysis plot as shown in Figure 26, perform exactly the same steps as in Instruction 4 and Instruction 5.

The Pan loci are now also selected in the **wgMLST** character experiment, in the form of a subset-based character view as described in 4.

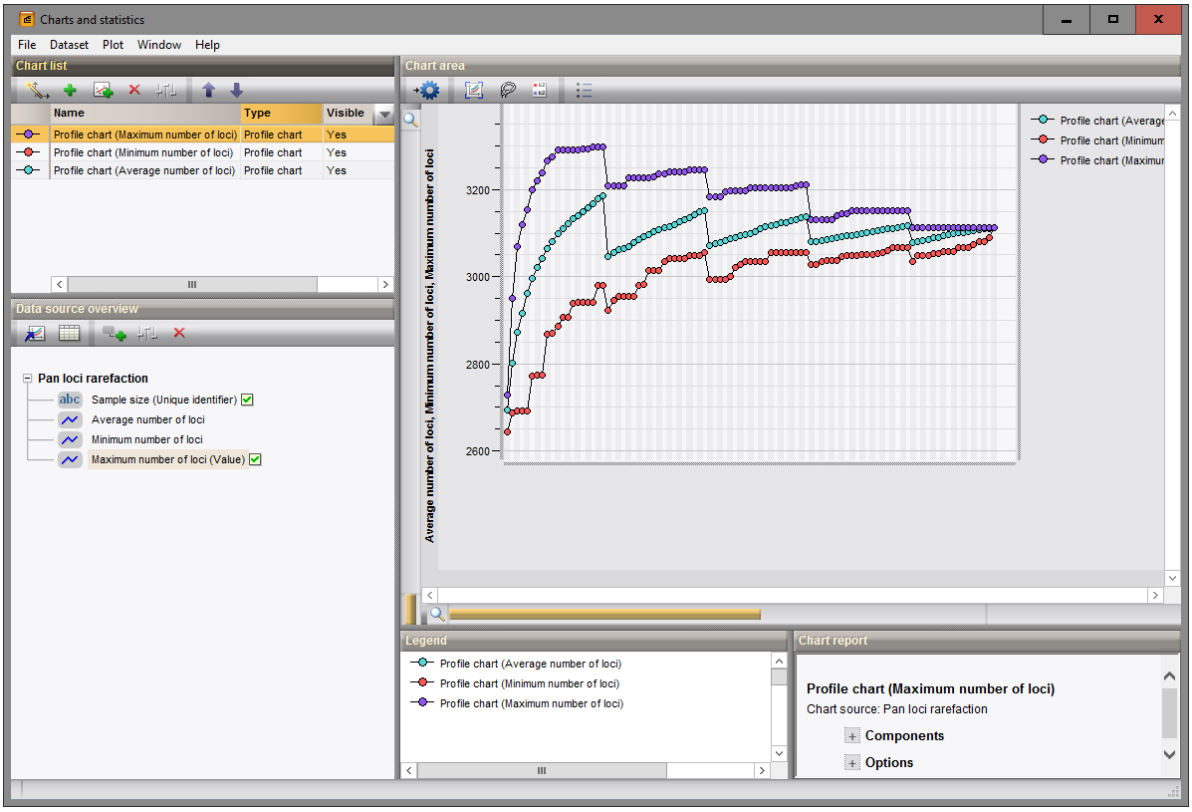

Figure 26: Pan locus analysis for all samples in the demonstration database (***Presence threshold***: 5%).

Bibliography

[1] David W Eyre, Tanya Golubchik, N Claire Gordon, Rory Bowden, Paolo Piazza, Elizabeth M Batty, Camilla LC Ip, Daniel J Wilson, Xavier Didelot, Lily O'Connor, et al. A pilot study of rapid benchtop sequencing of staphylococcus aureus and clostridium difficile for outbreak detection and surveillance. *BMJ open*, 2(3):e001124, 2012.

[2] Simon R Harris, Edward JP Cartwright, M Estée Török, Matthew TG Holden, Nicholas M Brown, Amanda L Ogilvy-Stuart, Matthew J Ellington, Michael A Quail, Stephen D Bentley, Julian Parkhill, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant staphylococcus aureus: a descriptive study. *The Lancet infectious diseases*, 13(2):130–136, 2013.

[3] Claudio U Köser, Matthew TG Holden, Matthew J Ellington, Edward JP Cartwright, Nicholas M Brown, Amanda L Ogilvy-Stuart, Li Yang Hsu, Claire Chewapreecha, Nicholas J Croucher, Simon R Harris, et al. Rapid whole-genome sequencing for investigation of a neonatal mrsa outbreak. *New England Journal of Medicine*, 366(24):2267–2275, 2012.