BioNumerics Tutorial:

Performing whole genome SNP analysis with mapping performed on the external calculation engine

1 Introduction

1.1 An introduction to whole genome SNP analysis

A Single Nucleotide Polymorphism (SNP) is a variation in a single nucleotide, which occurs at a specific position of the genome. SNPs are always defined with respect to a reference sequence. A SNP search or SNP analysis can therefore be regarded as a post-analysis on (aligned) sequences, in which SNPs are determined on one or more sample sequences, in relation to a reference sequence. When performed on whole genome sequences (WGS), this analysis is referred as **whole genome SNP (wgSNP) analysis**.

1.2 Whole genome SNP analysis in BioNumerics

This is a typical workflow for a wgSNP analysis in BioNumerics:

1.2.1 Choose a reference sequence

The choice of a reference sequence in a wgSNP analysis is very important, since only genomic information that is in common between the reference sequence and the sample sequence will be included in the analysis. With other words, any gene, integron, plasmid, etc. that is present in the reference but not in the sample (or vice-versa) will be left out. In order to obtain the highest possible resolution in a wgSNP analysis, the reference should be as similar as possible to the sample sequences. The reference sequence might be a closed, fully annotated genome sequence (e.g. downloaded from an online repository such as NCBI), but could as well be a de novo assembled sequence, consisting of multiple contigs (i.e. a draft genome).

1.2.2 Map sequence reads against the reference sequence

The most trivial way to ensure that genomic sequences are collinear (i.e. in the same frame and having the same length) for all isolates under investigation, is to map the trimmed sequence reads against the same reference sequence. This can be done locally on your desktop computer or using an external calculation engine.

1.2.3 Perform wgSNP analysis and filter out relevant SNPs

Each sample sequence, obtained via mapping to the reference sequence, is compared to this reference sequence and all base differences are recorded.

In addition to true point mutations, observed differences with the reference may be due to e.g. sequencing errors or larger indels and rearrangements. For phylogenetic analyses and strain typing it is therefore very important to retain only the relevant, high-quality SNPs. BioNumerics offers this functionality through various SNP filters. SNP filters are contained in a SNP template and their effect can be assessed in detail in the SNP filtering window, providing visual feedback and offering an easy link to the sequences and assemblies.

1.2.4 wgSNP clustering

A wgSNP clustering can be performed on the SNP matrix in the *Comparison* window.

2 Preparing the database

2.1 Introduction to the demonstration database

We provide a demonstration database for Staphylococcus aureus that contains NGS data (sequence read set data links), wgMLST results and wgSNP analyses. In this tutorial, we will proceed through all steps needed to perform a wgSNP analysis, and we will investigate the results obtained from a previous wgSNP analysis performed on isolates from a single study.

The demonstration database for Staphylococcus aureus can be downloaded directly from the *BioNumerics Startup* window, or restored from the back-up file available on our website:

2.2 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).

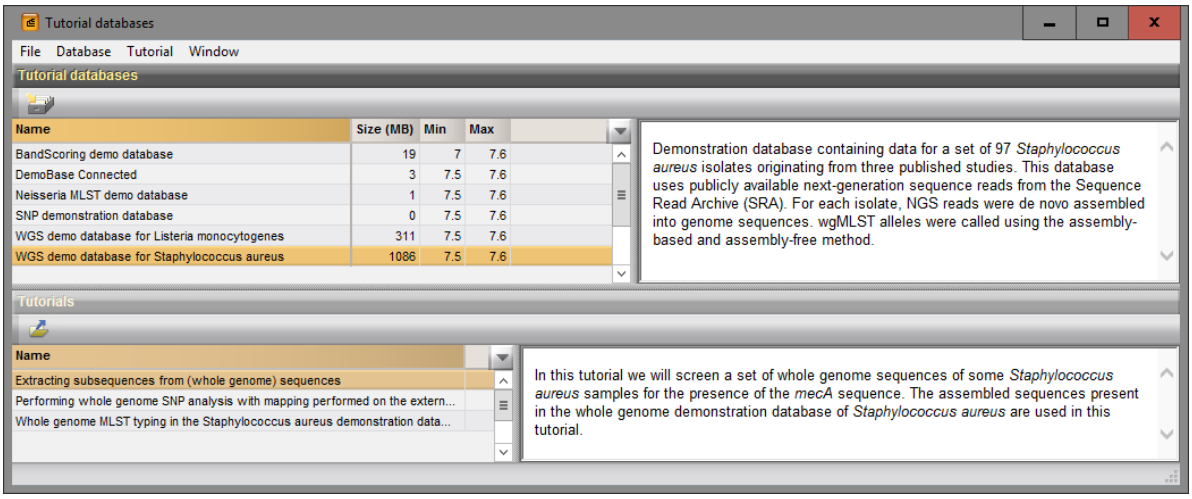

Figure 1: The *Tutorial databases* window, used to download the demonstration database.

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.

2.3 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).

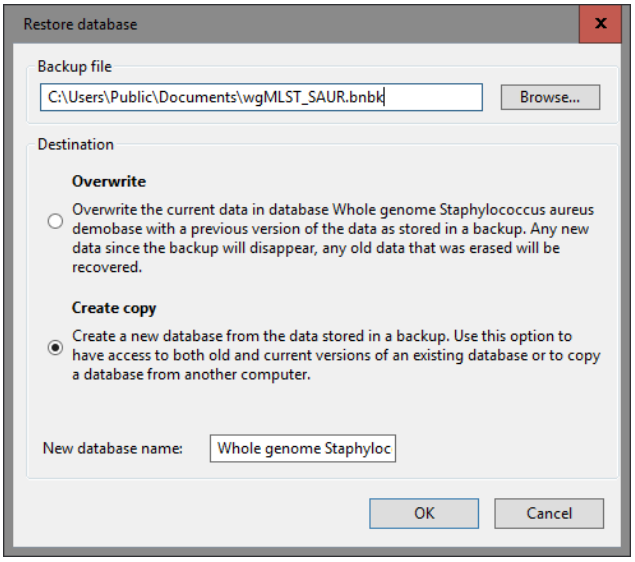

Figure 2: Restoring the whole genome demonstration database from the BioNumerics backup file `wg_SAUR.bnbk`.

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]). Sequence read set experiment type **wgs** contains the link to the sequence read set data on NCBI (SRA) with some raw data statistics (see Figure 3).

The full wgMLST analysis (de novo assembly, assembly-based calls and assembly-free calls) was performed on this set of samples using default settings and the *S. aureus* scheme prod_saur_v1_0. This data is used in the tutorial "wgMLST typing in the Staphylococcus aureus demonstration database". The results of the wgMLST analysis are stored in following experiments:

- 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 will be used in this tutorial to illustrate the *SNP filtering* window and the use of SNP filters (see 4.3).

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.

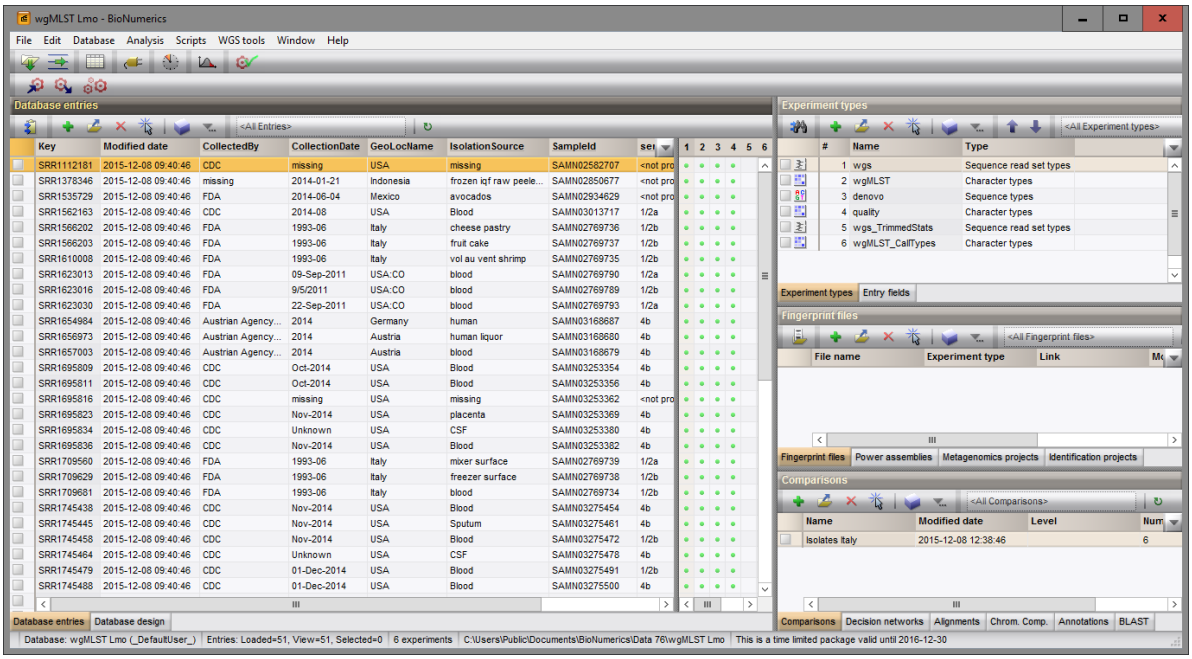

Figure 3: The *Main* window.

4 wgSNP analysis workflow in BioNumerics

4.1 Create a reference mapped sequence type

First, we will create a sequence type to store the reference mapped sequences in:

1. Click on the *Experiment types* panel to activate it and select ***Edit*** > ***Create new object...*** (). From the *Create a new experiment type* dialog box that pops up, select ***Sequence type*** and press <***OK***>.

2. In the *New sequence type* wizard, enter a ***Sequence type name*** (e.g. "My wgSNP") and press <***Next***> (see Figure 4).

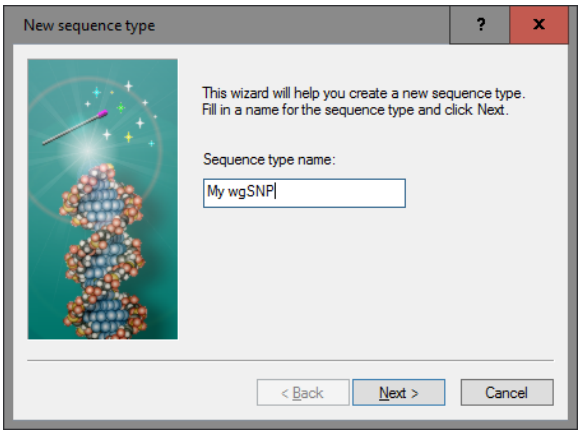

Figure 4: New sequence type experiment.

3. Leave the default ***Nucleic acid sequences*** option checked and check the option ***Use reference sequence as mapping template*** (see Figure 5).

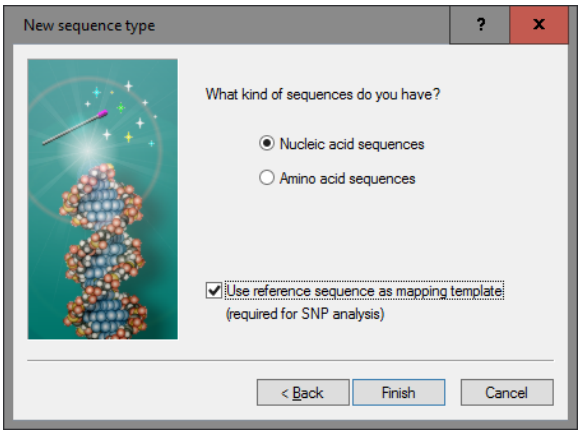

Figure 5: Use as mapping template.

4. Press <***Finish***> to create the reference mapped sequence type.

The first sequence that is imported in this sequence type will automatically be assigned as the reference. In this case, we will copy the sequence from the **denovo** experiment type of entry **ERR103401** to our new sequence type experiment:

5. Click on the dot in the *Experiment presence* panel that corresponds to the **denovo** experiment for entry **ERR103401**.

6. In the *Sequence editor* window that opens, select ***File*** > ***Save as...*** (**Ctrl+Shift+S**).

7. Highlight **My wgSNP** in the list and press <***OK***> (see Figure 6).

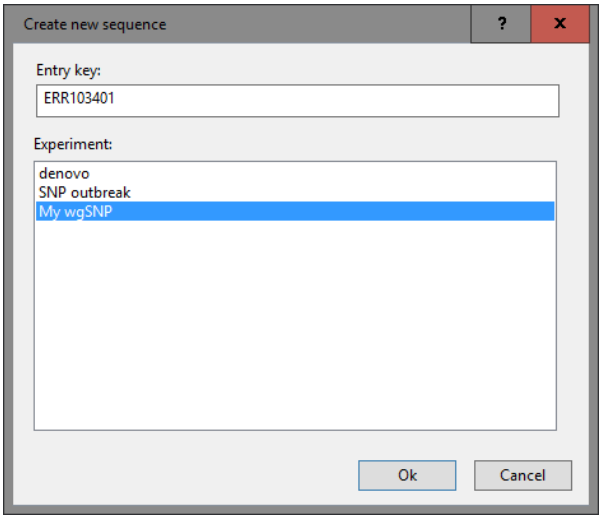

Figure 6: Save sequence to a new sequence type experiment.

8. Close the *Sequence editor* window.

We can check if this sequence is indeed used as reference:

9. In the *Experiment types* panel, double-click on **My wgSNP** to open its *Sequence type* window: the reference sequence is displayed here (see Figure 7).

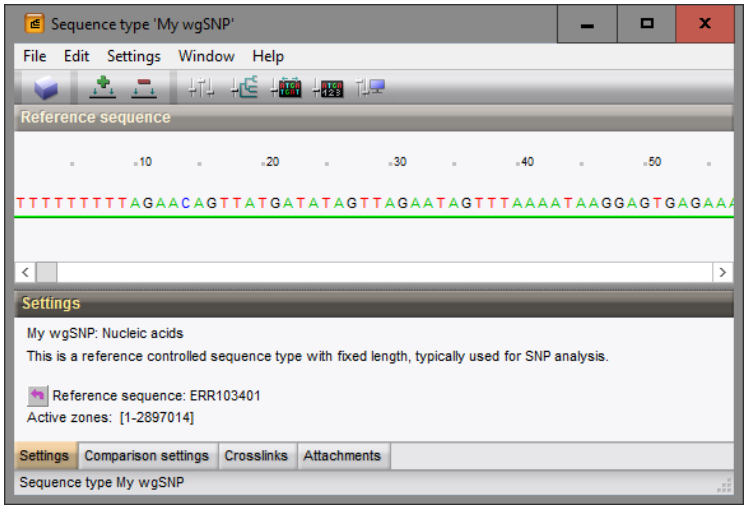

Figure 7: The new sequence type experiment and the reference sequence.

4.2 Map sequence reads against the reference sequence

Entries that we want to analyze using wgSNP need a sequence experiment that is obtained via mapping to the reference sequence, as this ensures collinearity of the sequences. For this, we will start from the corresponding sequence read sets.

4.2.1 Sequence read sets stored as links

Sequence reads can be imported as data links in BioNumerics using the ***Import sequence read set data as links*** import routine in the Import tree (***File*** > ***Import...*** (, **Ctrl+I**)). Importing sequence read sets as links and reference mapping calculations on the external calculation engine is only possible when the *WGS tools plugin* is installed in the BioNumerics database (***File*** > ***Install / remove plugins...*** ()). Installation of the plugin is only possible with a valid password and a project name, linked to a certain amount of credits. Please contact Applied Maths to obtain more information about the *WGS plugin* and the pricing.

In our demonstration database, the *WGS tools plugin* is already installed and a valid password and project name has been specified during installation. **Since no credits are assigned to the demo project, no reference mapping calculations can be performed on the external calculation engine in the demonstration database.** A reference mapping has already been calculated for the entries from the **Neonatal MRSA study** in our demonstration database, and the resulting sequences are used in 4.3.

For completeness, the steps to perform the reference mapping calculations on the external calculation engine are listed below:

10. Select the entries in the *Database entries* panel that you want to include in the SNP analysis.

11. Select ***WGS tools*** > ***Submit jobs...*** () to display the *Submit jobs* dialog box.

12. Make sure the correct sequence type experiment is selected as ***Reference mapping*** and uncheck all other jobs (see Figure 8).

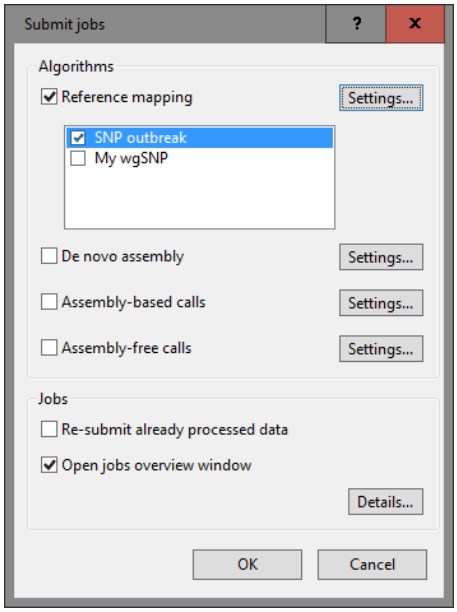

Figure 8: Submit jobs.

13. Press the <***Settings***> button to choose on of the two available algorithms (***Bowtie*** or ***Applied Maths mapper***) and verify the algorithm settings (see Figure 9). Press <***OK***> to close the *Reference mapping settings* dialog box again.

14. Back in the *Submit jobs* dialog box, click the <***Details***> button to get some detailed information on the credits needed to post the selected jobs (see Figure 10).

15. Close the *Details* dialog and press <***OK***> to launch the jobs.

The *Calculation engine overview* window will open, in which the status of the jobs can be followed (see Figure 11).

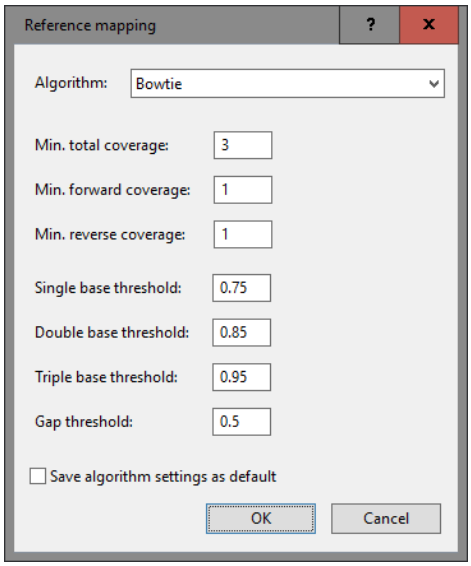

Figure 9: Mapping algorithms.

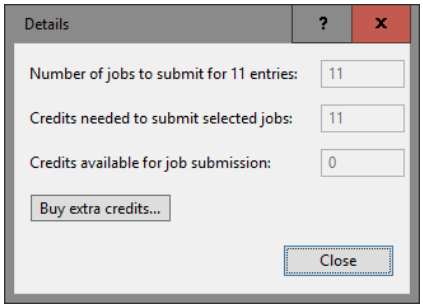

Figure 10: Details on credits needed and available credits.

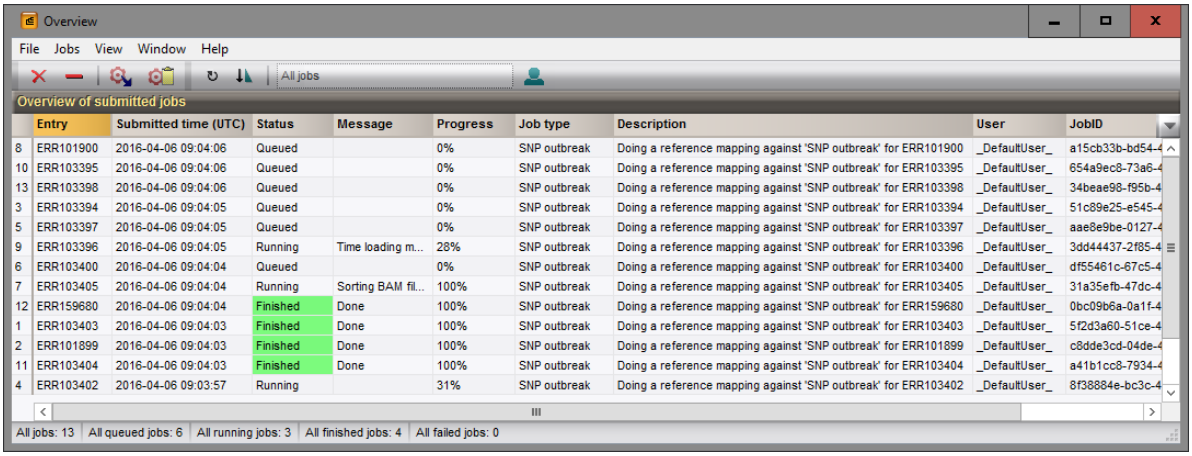

Figure 11: Job overview.

16. Close the *Calculation engine overview* window.

When insufficient credits are available, an error message will appear (see Figure 12). Since no credits are assigned to the demo project, this error message will pop up when following this workflow in the demonstration database when entries are selected for which no reference mapping is present. Please consult Applied Maths for more information about the purchase of credits.

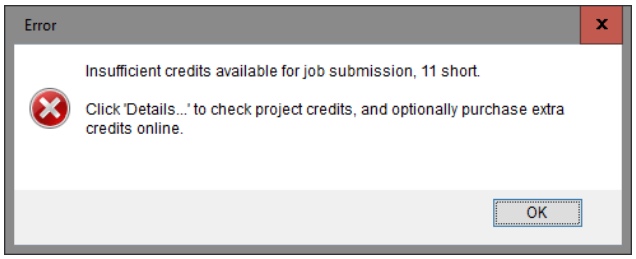

Figure 12: Insufficient credits available.

When a reference mapping is already present for the submitted entries (and the ***Re-submit already processed data***) option was unchecked in the *Submit jobs* dialog box, an information message will appear, saying that no jobs are submitted to the calculation engine (see Figure 13). When the selection in the demonstration database only contains entries from the **Neonatal MRSA study**, this information message will pop up.

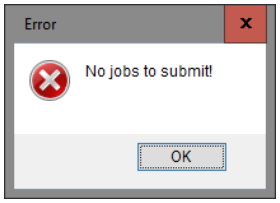

Figure 13: No jobs to submit.

4.2.2 Sequence read sets stored inside the database

Sequence read sets can also be stored inside the database with the ***Import sequence read set files*** import routine in the Import tree (***File*** > ***Import...*** (, **Ctrl+I**)). One disadvantage of this option is the storage size of the sequence read sets.

The steps to perform the reference mapping calculations on the external calculation engine for sequence read sets that are stored inside the database are the same as for the sequence read sets for which the links are stored inside the database (see 4.2.1).

4.3 Perform wgSNP analysis and filter out revelant SNPs

In our demonstration database, a reference mapping has already been calculated for the entries from the **Neonatal MRSA study**. The resulting sequences are stored in the **SNP outbreak** sequence type and will be used in this section to start a SNP analysis and to illustrate the use of SNP filters in the *SNP filtering* window.

17. Make sure no selection is present in the *Main* window by pressing the **F4**-key.

18. In the *Database entries* panel, select the **Neonatal MRSA study** view from the list (see Figure 14) and use ***Edit*** > ***Select all*** (**Ctrl+A**) to select all 14 entries contained in this study.

19. Select ***Analysis*** > ***Sequence types*** > ***Start SNP analysis...*** to start the *SNP analysis* wizard.

20. Select **SNP outbreak** as ***Experiment type*** and press <***Next***> (see Figure 15).

A number of predefined SNP templates are available.

21. Highlight the ***Strict filtering*** template and press <***Next***> (see Figure 16).

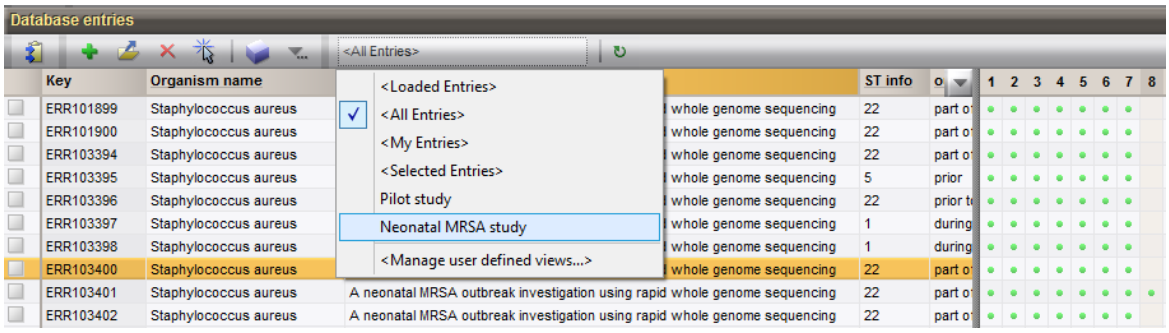

Figure 14: Select a view from the list.

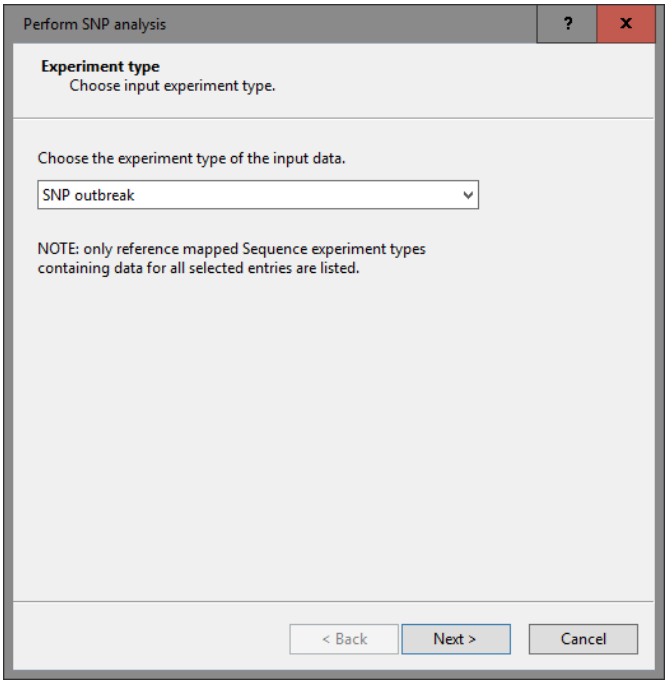

Figure 15: Select the input sequences.

22. Check ***Open SNP analysis window*** and press <***Finish***> (see Figure 17).

It will take a few moments to load the sequences and apply the filters from the SNP template. The resulting *SNP filtering* window is shown in Figure 18.

This window consists of following panels:

- The *Entries* panel shows all entries that are included in the SNP analysis, with all entry information fields. Two additional fields are present: 'Total' shows the raw number of SNPs (i.e. without any SNP filter applied) and 'Retained' shows the number of SNPs after applying all active SNP filters for the sample sequence.
- The *Filters* panel shows the list of SNP filters that are applied, with the 'Info' column showing additional information regarding the filter and applied settings (if applicable). This list is initially populated from the SNP template, but SNP filters can be added or removed and their settings can be changed.
- The *SNP Positions* panel shows information on all positions where at least one SNP was detected. For each SNP filter that is listed in the *Filters* panel, a column is displayed with the filter's result on each

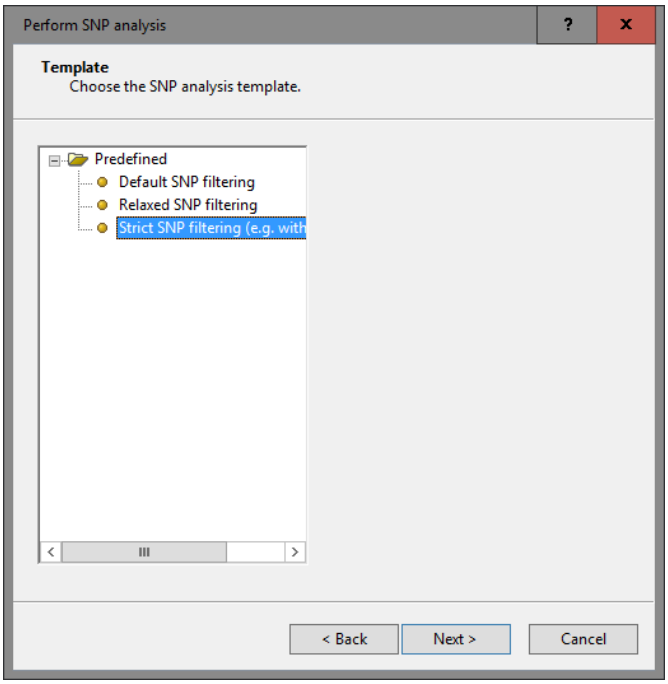

Figure 16: Choose a SNP filtering.

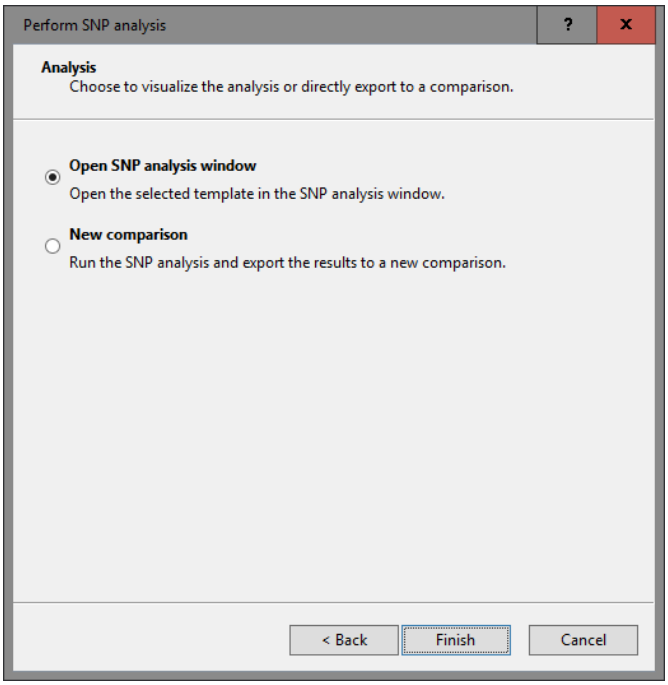

Figure 17: Open SNP analysis window.

position. The bottom of this panel shows a sub-panel with the details on the highlighted position, i.e. showing the base and filter results for all the sample sequences on that position.

- The *Entry SNPs* panel lists the SNPs for the highlighted entry in the *Entries* panel.
- The *Genome* panel shows the SNPs on a genome view.
- The *Tracks* panel in default view is displayed as a tab with the *Entries* panel. With this panel, you can determine which tracks are plotted in the *Genome* panel.

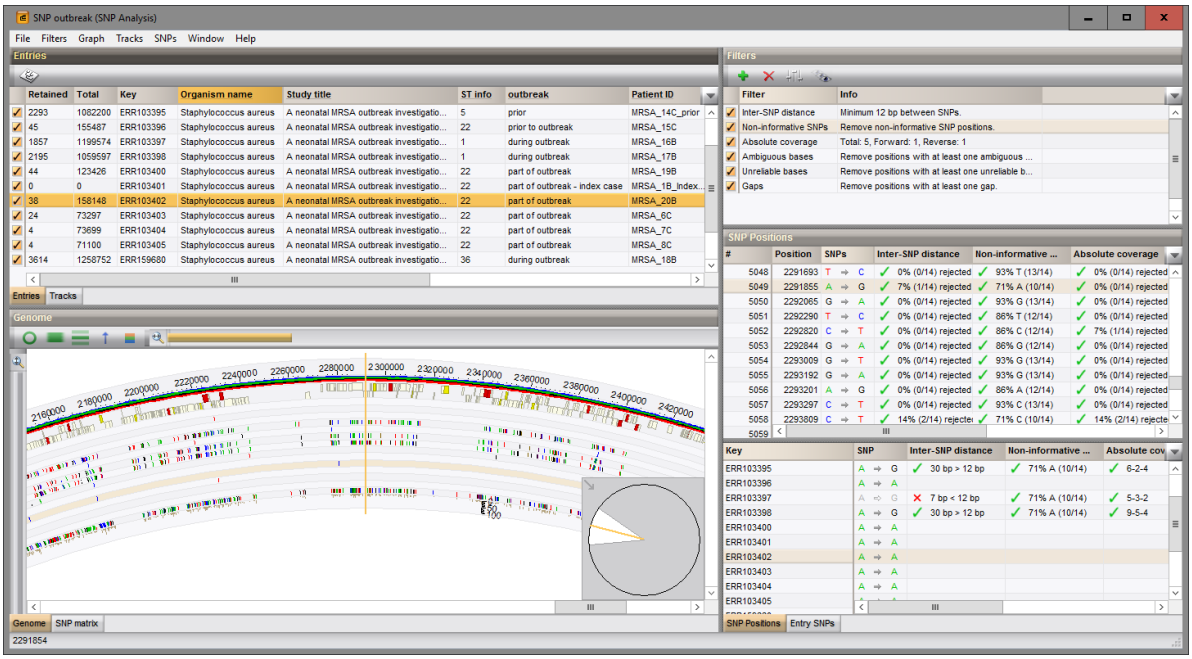

Figure 18: SNP analysis window.

- The *SNP matrix* panel shows the resulting SNP matrix, as it would be exported.

Whenever possible, the cursor position is synchronized between the different panels:

23. Click on a position in the *SNP Positions* panel for example.

The details in the bottom part of the panel are updated and so is the *Genome* panel: the graph will show the position. Furthermore, the clicked position in the *SNP Positions* panel will appear highlighted in the *Entry SNPs* panel, *only* if the currently highlighted entry in the *Entries* panel has a SNP at that position.

24. Double-click a position in the details panel (bottom part of the *SNP Positions* panel) or in the *Entry SNPs* panel.

This action will open the *Sequence editor* window of the corresponding sequence, with this position highlighted. If a sequence assembly is available in BioNumerics, the will be active and selecting ***File*** > ***Open assembler*** ( ) will open the assembly.

Sequence assemblies are not available when the remapping was performed on the calculation engine.

25. A SNP filter can be added with ***Filters*** > ***Add filter...*** ( ).

26. Check or uncheck an individual SNP filter in the *Filters* panel to view its effect.

When the toggle ***Filters*** > ***Toggle rejected SNP visibility*** is unchecked ( ), the positions in the *SNP Positions* panel and the *Entry SNPs* panel will be limited to the retained SNPs, i.e. those SNPs that have passed the applied SNP filters.

When the toggle is checked ( ) the listed positions in both panels correspond to the total (i.e., unfiltered) SNP set.

27. Click on the tab of the *SNP matrix* panel to show the SNP matrix (see Figure 19).

28. Select ***File*** > ***Export to comparison...*** ( ) to export the SNP matrix to a comparison.

In the *Comparison* window a cluster analysis can be calculated based on the exported SNP data (see 5.1).

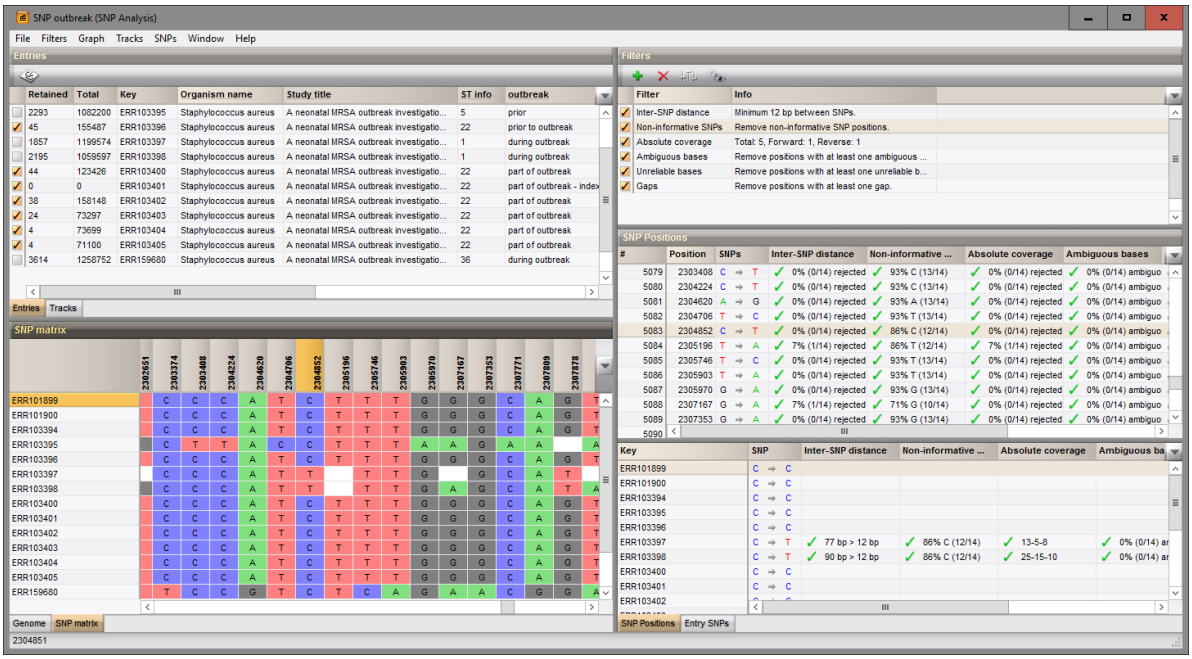

Figure 19: SNP matrix displayed.

5 Follow-up analysis

5.1 Cluster analysis on SNP data

1. Selecting ***File*** > ***Export to comparison...*** ( ) in the *SNP filtering* window exports the SNP matrix to a new comparison (see 4.3).

In this comparison, the SNP matrix is available as a *character aspect* of the **SNP outbreak** sequence experiment type (see Figure 20).

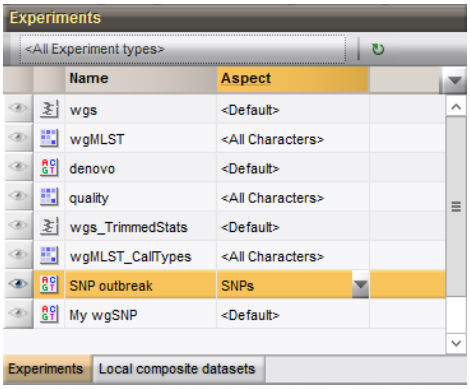

Figure 20: SNPs character aspect in the *Comparison* window.

We can now create a cluster analysis based on the SNP data, in the same way that a similarity-based clustering is performed in BioNumerics:

2. Make sure the **SNP outbreak** experiment is selected in the *Experiments* panel and select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***.

Only multi-state similarity coefficients are suitable for clustering of SNP data:

3. Select e.g. the ***Categorical (values)*** coefficient and press <***Next***> (see Figure 21).

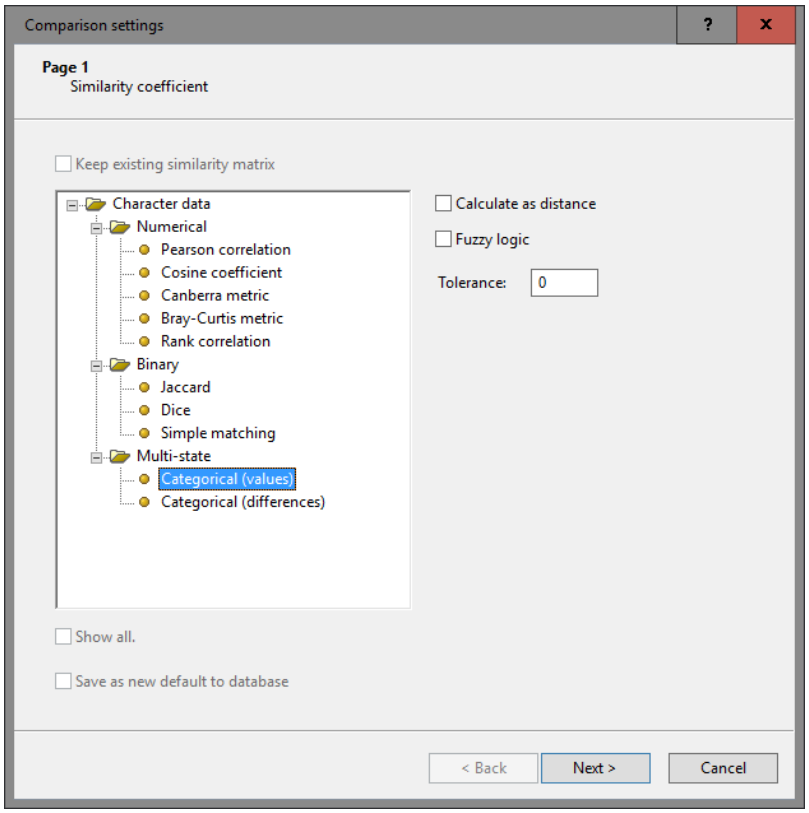

Figure 21: The categorical similarity coefficient.

4. Leave the default settings enabled and press <***Finish***> to calculate the dendrogram.

The resulting dendrogram is displayed in the *Dendrogram* panel of the *Comparison* window (see Figure 22).

5. To clear the selection, press the **F4**-key.

The dendrogram contains one well-defined cluster at 99% and a number of unrelated strains.

6. Hold the **CTRL**-key and click on this cluster of related strains to select the 10 entries in the database.

This cluster contains all strains with MLST sequence type 22 (see info field '"ST info") and corresponds to the strains that were identified as part of the outbreak in the published study (see Figure 22).

7. Save the comparison with the dendrogram by selecting ***File*** > ***Save as...***. Specify a name (e.g. **Neonatal study**) and press <***OK***>.

5.2 Comparison with wgMLST results

For comparison purposes, we will create a cluster analysis for the same entries, but now based on wgMLST data:

8. Click on the icon left of **wgMLST** and select the ***wgMLST loci*** aspect from the drop-down list to visualize the wgMLST allele numbers in the *Experiment data* panel.

9. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***, highlight the ***Categorical (values)*** coefficient and press <***Next***>.

10. Leave the default settings enabled and press <***Finish***> to calculate the dendrogram.

The resulting dendrogram is displayed in the *Dendrogram* panel of the *Comparison* window (see Figure 23).

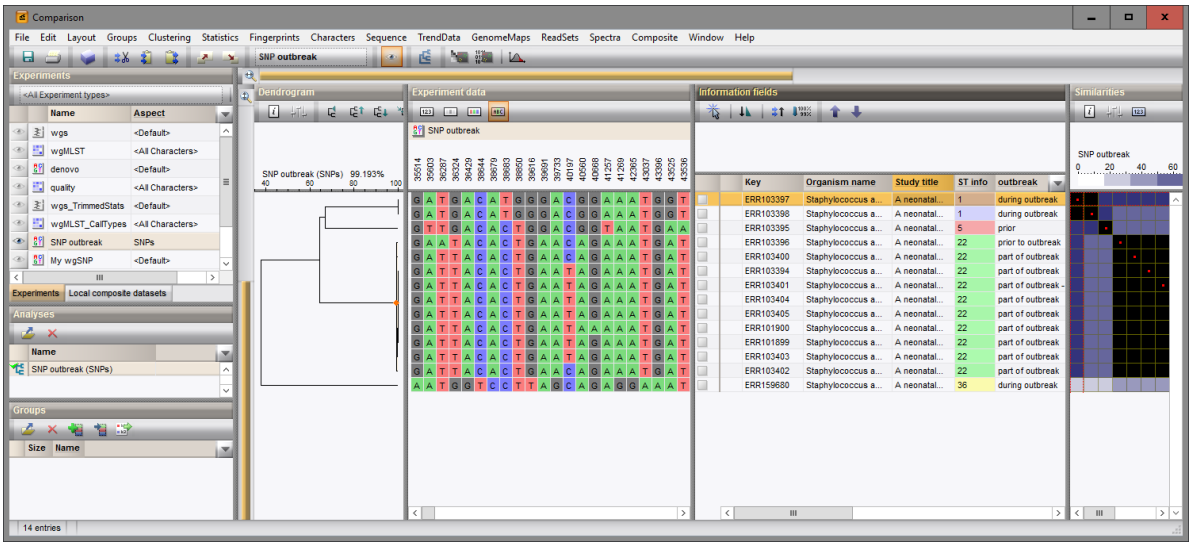

Figure 22: The *Comparison* window: wgSNP dendrogram.

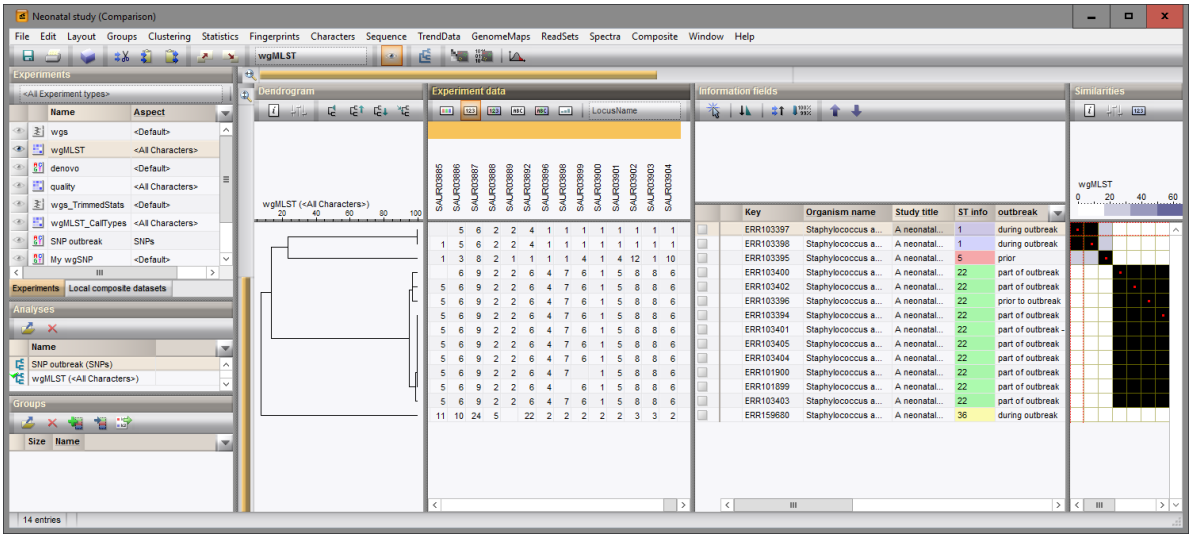

Figure 23: The *Comparison* window: wgMLST dendrogram.

Via the *Analyses* panel, you can switch back and forth between the two dendrograms. From this visual inspection, it is clear that both analyses correlate very well.

11. Select ***Clustering*** > ***Congruence of experiments...***.

Both experiments show a very high congruence of 94%.

5.3 Zooming in on the outbreak

As an alternative to the procedure described in 4.3, a SNP filtering can also be started from the *Comparison* window. We will illustrate this workflow by ”zooming in” on the isolates that belong to the outbreak:

12. First, make sure no entries are selected in the *Comparison* window by pressing **F4**.

13. Select the 10 entries that belong to the largest cluster in the dendrogram, e.g. using **Ctrl+click** on the corresponding branch in the dendrogram.

14. Return to the *Main* window to create a new comparison for the selected entries via ***Edit*** > ***Create new object...*** () or directly using the **Alt+C** keyboard shortcut.

A new *Comparison* window pops up with the ten entries that are associated with the outbreak.

15. Select ***File*** > ***Save*** (, **Ctrl+S**), enter e.g. **Outbreak** as name and press <***OK***>.

16. Click on the  icon left of **SNP outbreak** to visualize the sequences in the *Experiment data* panel.

17. Select ***Sequence*** > ***Open SNP window...***

18. Highlight a SNP template from the *SNP analysis* wizard, e.g. ***Strict filtering*** (the same as we used previously).

This action shows the *SNP filtering* window again, as discussed in 4.3.

19. Export the SNP matrix back to the comparison via ***File*** > ***Export to comparison...*** ().

This time, a ***SNP data name*** will be prompted for. Saving SNP matrices under different names will create multiple character aspects for the same sequence type. These aspects are available via the 'Aspect' drop-down list in the *Experiments* panel and can each be used to create cluster analyses from.

20. Enter e.g. **SNPs strict** and press <***OK***> in the dialog.

21. Return to the **Outbreak** comparison, where the SNP matrix is now displayed.

We can again calculate a dendrogram:

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

23. This time, select the ***Categorical (differences)*** coefficient, specify a ***Scaling factor*** of "1" and press <***Next***>.

24. Check ***Single linkage*** for ***Method*** and press <***Finish***> to calculate the dendrogram.

The resulting dendrogram indicates a sub-structure in the isolates that are associated with the outbreak: two well-defined groups are found (see Figure 24).

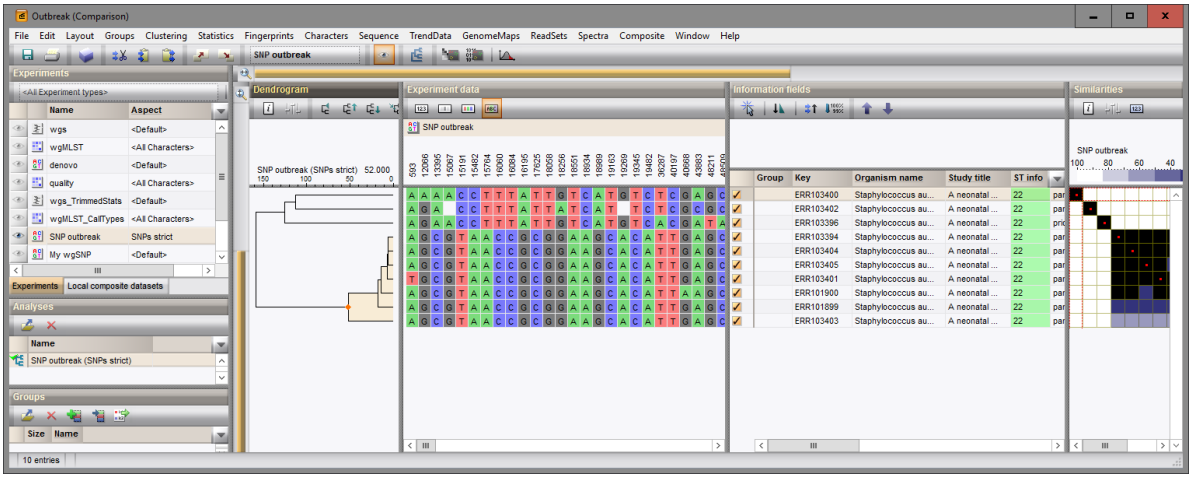

Figure 24: Zooming in on the outbreak: two groups.

5.4 Exporting SNP data

If needed, SNP data can be exported as a character set. We will illustrate this for the **SNP strict** aspect of **SNP outbreak**:

25. In the *Comparison* window, select **SNP strict** from the 'Aspect' drop-down list next to **SNP outbreak**.

26. Select ***File*** $>$ ***Export*** $>$ ***Export character data...***.

27. In the *Export character data* dialog box, make sure ***Export mapped values*** is checked and press <***OK***>.

The exported SNP matrix will open automatically in MS Excel.

Alternatively, the data displayed in the *SNP matrix* panel of the *SNP filtering* window can be exported using the column properties button () and selecting e.g. ***Save content to file***.

Other applications might require the list of SNPs per entry formatted as (pseudo-)sequence:

28. In the *Comparison* window, with the **SNP Strict** aspect still selected, use ***File*** $>$ ***Export*** $>$ ***Export sequences (fasta)...***.

The `export.txt` file that opens is a multi-FASTA file with each row of the SNP matrix represented by a sequence.

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.