BioNumerics Tutorial:

Performing whole genome SNP analysis with mapping performed locally

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 Example data

Example data that will be used in this tutorial can be downloaded from the Applied Maths website: `http://www.applied-maths.com/download/sample-data`, "FASTQ files"). The data set (see Figure 1) contains:

- 10 gzipped fastq files of 5 paired end read data file pairs coming from *Staphylococcus aureus*.
- An Excel file `Strain information.xlsx` containing some meta data on the sequence read sets.
- A sequence stored in a text file (`Reference.txt`) that will be used as reference sequence in the mapping step.

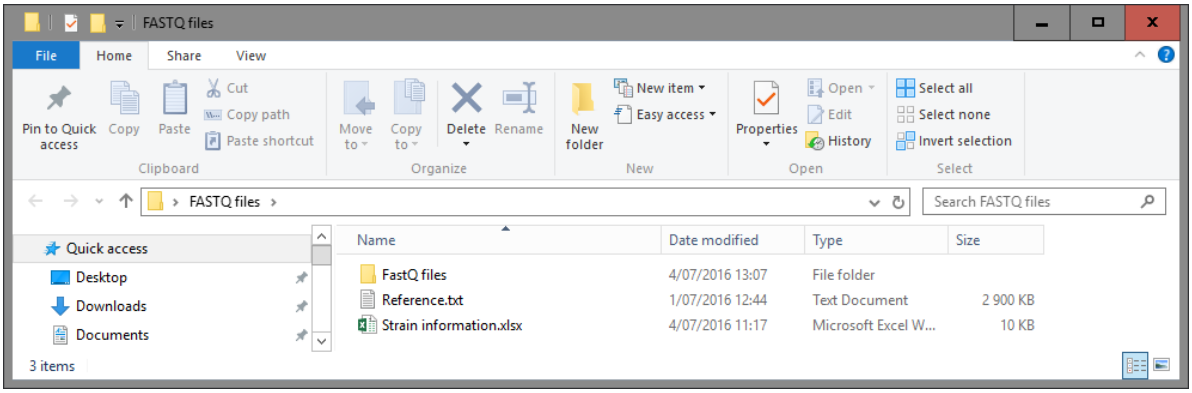


Figure 1: Example data.

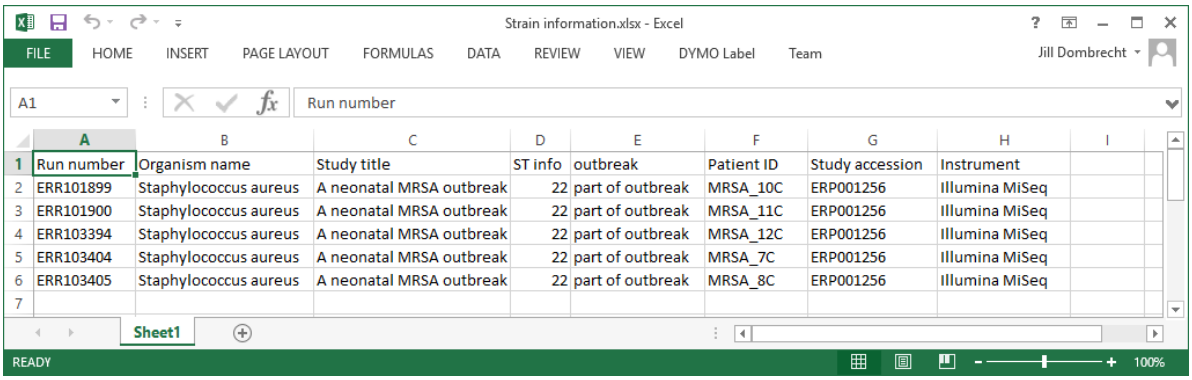
2.2 Importing strain information

In this part, we will import the metadata on the sequence read sets from the external "Strain information.xlsx" Excel file (see Figure 2).

1. Create a new database (see tutorial "Creating a new database") or open an existing database.
2. Select ***File*** > ***Import...*** (, **Ctrl+I**) to open the import wizard.
3. Choose the option ***Import fields (Excel file)*** under the ***Entry information data*** in the tree (see Figure 3) and click <***Import***>.
4. A dialog box pops up, allowing you to browse for an Excel file as source. Press <***Browse***>, navigate to the "Strain information.xlsx" file saved to your computer, and press <***Next***>.

The next dialog box allows you to set import rules. For each import source (i.e. Excel column), a database destination can be specified.

5. Make a multiple selection for all rows. Do this by selecting the first row (**Run number**) and while holding the **Shift**-key, double-click on the last row (**Instrument**). Select <***Edit destination***>, select "Entry info field" as destination and click <***OK***>.



Run number	Organism name	Study title	ST info	outbreak	Patient ID	Study accession	Instrument
ERR101899	Staphylococcus aureus	A neonatal MRSA outbreak	22	part of outbreak	MRSA_10C	ERP001256	Illumina MiSeq
ERR101900	Staphylococcus aureus	A neonatal MRSA outbreak	22	part of outbreak	MRSA_11C	ERP001256	Illumina MiSeq
ERR103394	Staphylococcus aureus	A neonatal MRSA outbreak	22	part of outbreak	MRSA_12C	ERP001256	Illumina MiSeq
ERR103404	Staphylococcus aureus	A neonatal MRSA outbreak	22	part of outbreak	MRSA_7C	ERP001256	Illumina MiSeq
ERR103405	Staphylococcus aureus	A neonatal MRSA outbreak	22	part of outbreak	MRSA_8C	ERP001256	Illumina MiSeq

Figure 2: Run information stored in an Excel file.

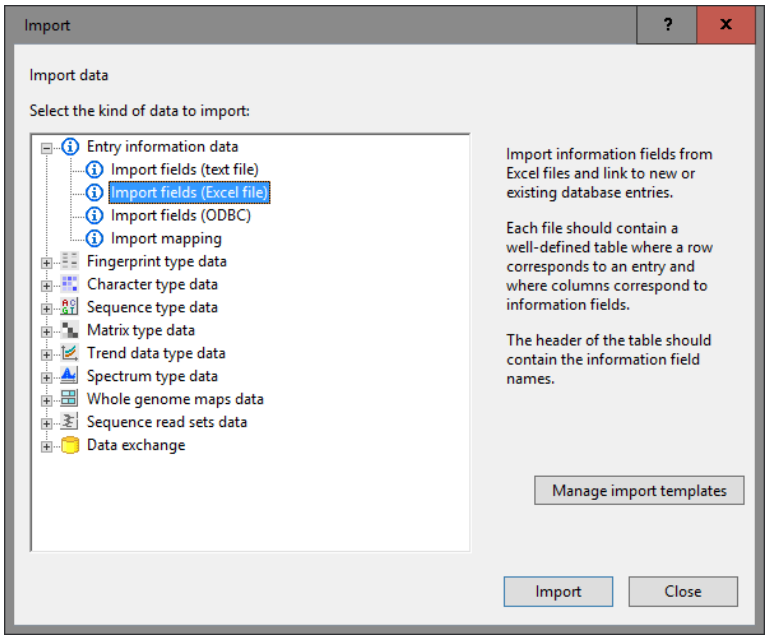


Figure 3: Import tree.

Optionally you can change the names of the new information fields to store the information in.

6. Click <***OK***> and click <***Yes***> to confirm the creation of the new information fields.

The import rules are updated in the grid (see Figure 4).

7. Optionally, you can do a preview of what you are about to import. Press <***Preview...***> to open the preview.

8. Click <***Next***> and <***Finish***> to finish the creation of the import template for the database information fields.

9. Enter a meaningful name (and optionally a description) for the created import template e.g. “Import of run information”, and click <***OK***>.

10. Then choose the newly created import template from the list and click <***Next***>.

11. The next dialog allows you to confirm the creation of the 5 new entries in the database. Click <***Finish***>.

After the import, 5 new entries have been added to the database (see Figure 5).

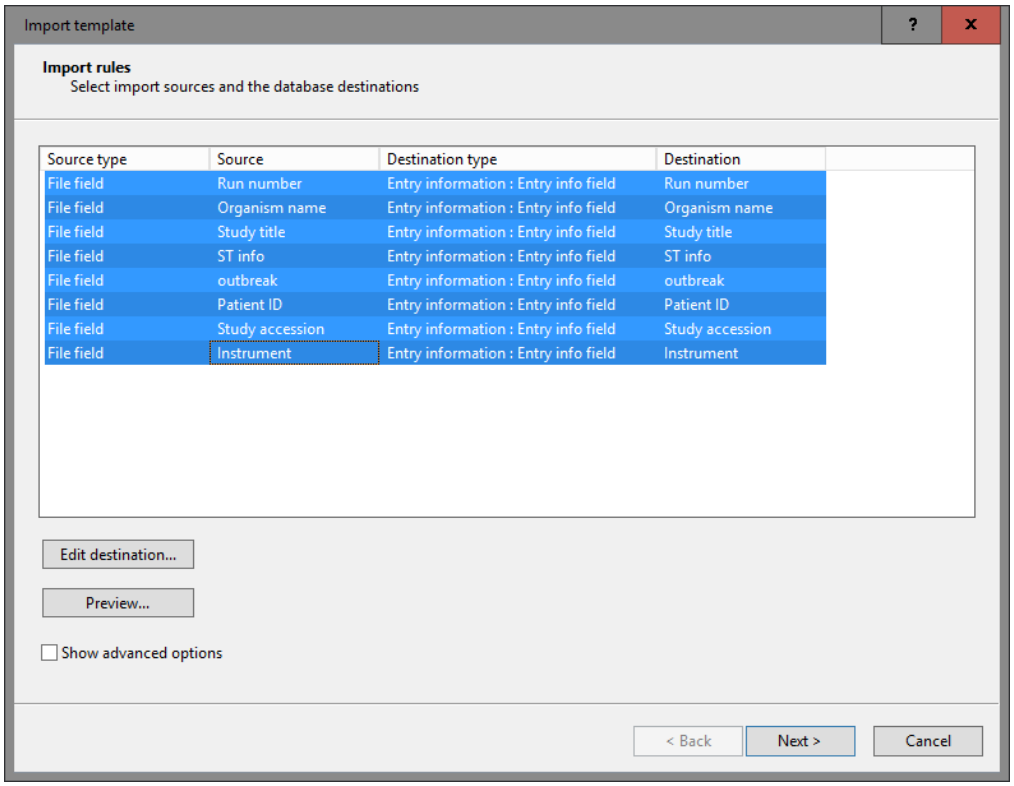


Figure 4: Import rules.

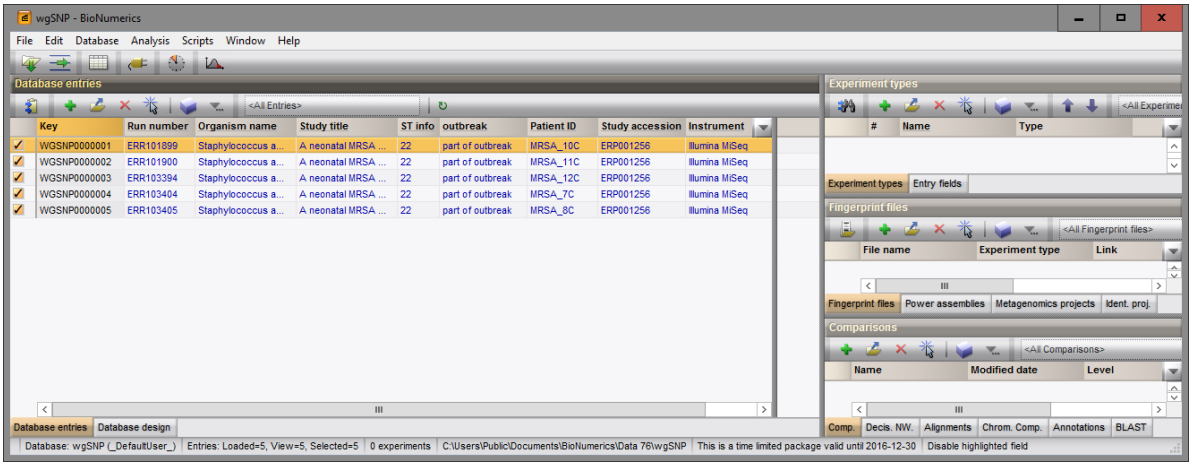


Figure 5: The *Main* window after import of the meta data.

2.3 Importing sequence read sets

Once the metadata of the different samples is imported, the actual sequence read sets can be linked to these entries in the database.

12. Select ***File*** > ***Import...*** (, **Ctrl+I**) to open the *Import* dialog box again.

13. Select the option ***Import sequence read set files*** under ***Sequence read sets data***, and press <***Import***> to start the *Import sequence read sets* wizard (see Figure 6).

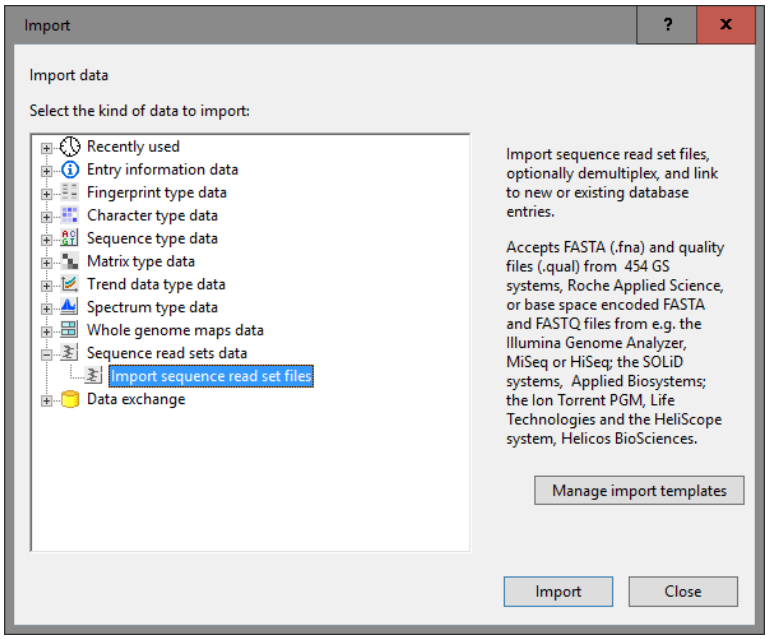


Figure 6: Import sequence read sets.

Sequence reads can also be imported as data links in BioNumerics using the ***Import sequence read set data as links*** import routine in the Import tree. Importing sequence read sets as links 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 the first step of the wizard you need to browse for the sequence reads set files containing the data.

14. Press <***Browse***>, navigate to the correct location, select all 10 files in the `FastQ files` folder and press <***Open***> to add the selected files to the import dialog. Press <***Next***> to proceed.

15. No demultiplexing is needed so press <***Next***> to continue.

Now you need to define how the data should be stored in the database. The default template **Example import** will only retain the SRA run accession numbers from the file names and store this information in the BioNumerics ***Key*** field. In this example, the accession numbers need to be linked to the ***Run number*** column, so a new import template needs to be created:

16. Press the <***Create new***> button to call the ***Import rules*** dialog.

17. In the ***Import rules*** dialog, select the only row titled ***File*** and click <***Edit destination***>.

The files names contain the information stored in the ***Run number*** column in our BioNumerics database. We will link the file names to the ***Run number*** entry information field.

18. Select the entry info field ***Run number*** as the data destination (see Figure 7) and press <***OK***>.

19. Click <***Preview***> to inspect the result of the import rule.

It is clear that the part **_1.fastq** still has to be removed to get the SRA run accession number.

20. Click <***Close***> to close the preview window.

21. Check the box next to ***Show advanced options*** and then click <***Edit parsing***>.

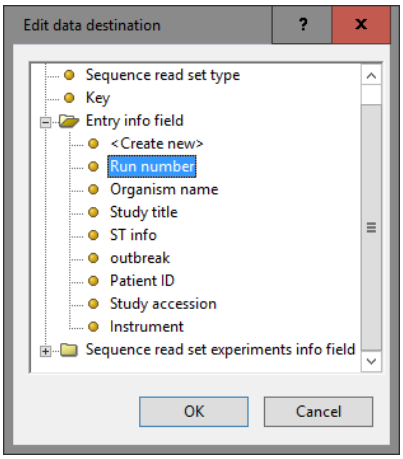


Figure 7: Link to Run number column.

22. In the *Edit data destination* dialog box, fill in "[DATA]_*" as ***Data parsing string*** (see Figure 8). Click <***Preview***> to double-check that the output now corresponds to the SRA run accession number. Click <***OK***>.

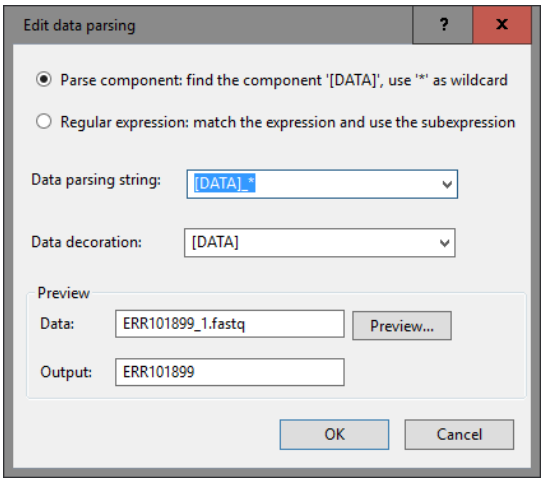


Figure 8: Parsing string.

The import rule is updated in the grid.

23. Click <***Preview***> to inspect the result of the import rule.

24. Close the preview and click <***Next***> to proceed.

25. Check the field ***Run number*** to use it as link field (see Figure 9), and press <***Finish***> to complete the import template.

26. Enter a meaningful name, such as "Import of FASTQ files". Optionally, enter a description for the created import template, and click <***OK***>.

27. Select the import template from the list, and select ***Create new*** from the ***Experiment type*** list (see Figure 10).

28. Click <***Next***> and specify a sequence type name, e.g. "wgs".

29. Click <***OK***> and confirm the creation of the experiment.

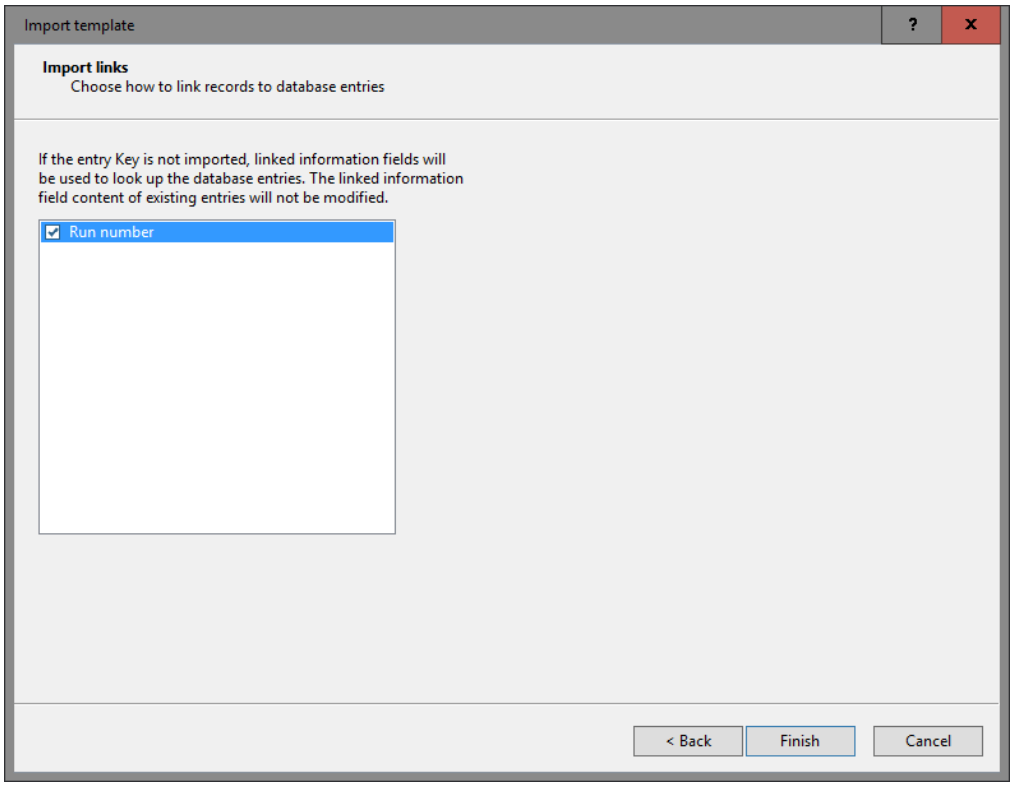


Figure 9: Link field.

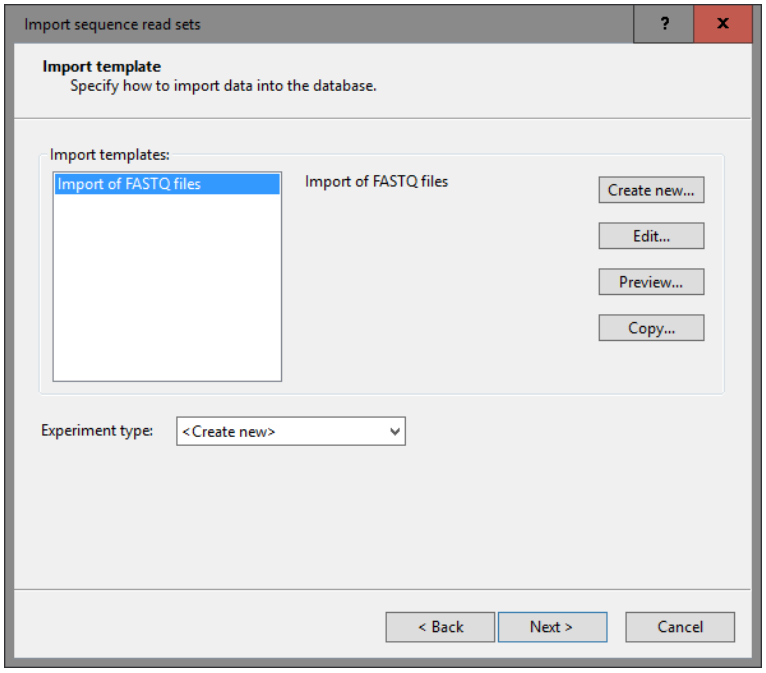


Figure 10: Import template.

30. Leave the option **update 5 entries** checked. Press <***Finish***> to start the import of the sequence read sets.

The sequences are linked to the existing entries in the database. Once the import is completed, select one of the green experiment dots from the *Experiment presence* panel to visualize some basic statistics on the

imported sequence read sets.

31. Close the sequence read set card.

3 wgSNP analysis workflow in BioNumerics

3.1 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 11).

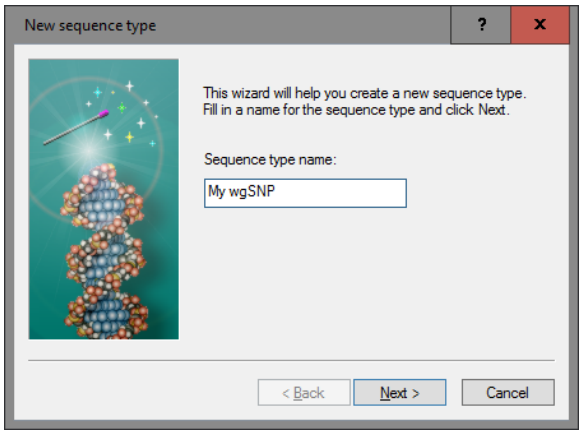


Figure 11: New sequence type experiment.

3. Leave the default ***Nucleic acid sequences*** option checked and check the option ***Use reference sequence as mapping template*** (see Figure 12).

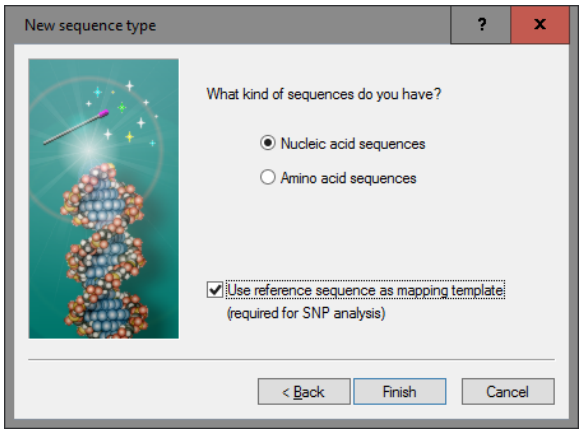


Figure 12: Use as mapping template.

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

3.2 Import a reference sequence

The first sequence that is imported in the newly created reference mapped sequence type will automatically be assigned as the reference. Here, we will import the sequence from the text file `Reference.txt` (see Figure 13). This sequence corresponds to the denovo assembled genome sequence of strain with run number **ERR103401**.

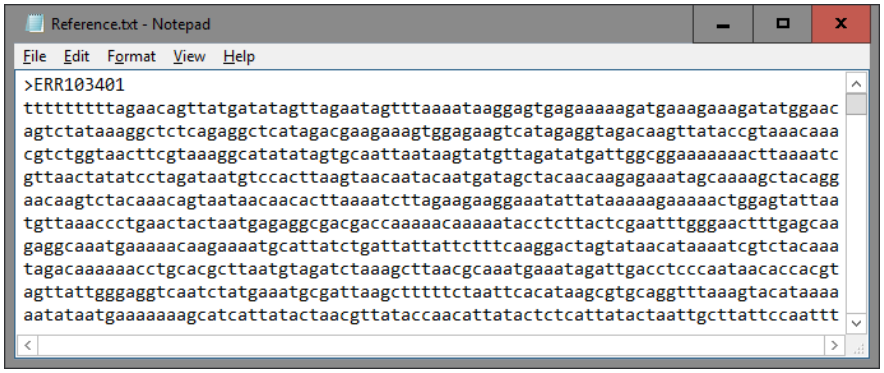


Figure 13: Reference sequence.

5. Select ***File*** > ***Import...*** (, **Ctrl+I**) to open the import wizard.

6. Choose the option ***Import FASTA sequences from text files*** under the ***Sequence type data*** in the tree (see Figure 14) and click <***Import***>.

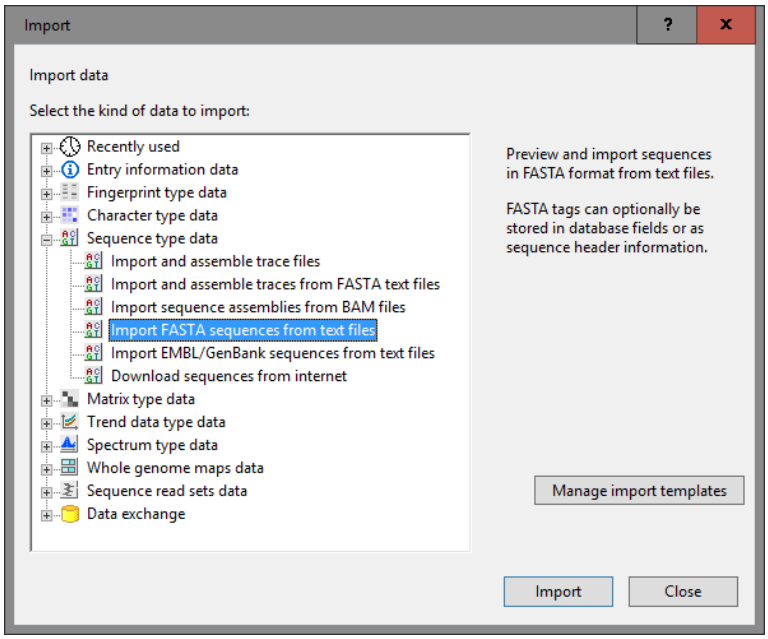


Figure 14: The Import tree.

7. Browse for the `Reference.txt` file, press <***Open***> and press <***Next***> twice.

8. Press <***Create new***> to define a new import template.

The first (and only) FASTA field in the `Reference.txt` file corresponds to the ***Run number***. We will link this field to the ***Run number*** field in our BioNumerics database.

9. Double-click on **Field 1** and change the destination to ***Run number*** (see Figure 15) and press <***OK***>.

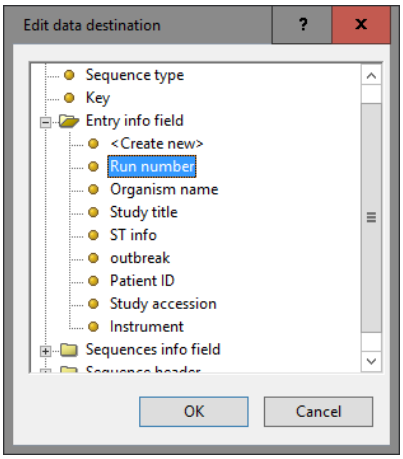


Figure 15: Link to the ***Run number*** field.

10. Press <***Next***> and <***Finish***>.

11. Enter a meaningful template name, such as "Import of reference sequence". Optionally, enter a description for the created import template, and click <***OK***>.

12. Make sure the newly created template is selected together with the ***My wgSNP*** sequence type (see Figure 16) and press <***Next***>.

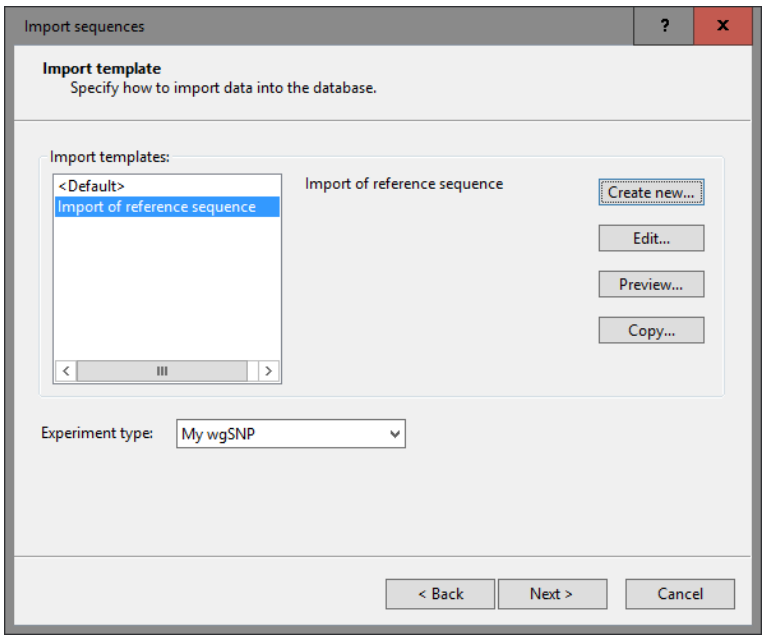


Figure 16: Import template.

13. Press <***Finish***>.

A new entry is added to the database and the imported sequence is stored in the ***My wgSNP*** sequence type: a green colored dot is available in the second column for this new entry.

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

14. In the *Experiment types* panel, double-click on **My wgSNP** to open its *Sequence type* window: the reference sequence is displayed in the *Settings* panel.

15. Close the *Sequence type* window.

3.3 Map sequence reads against the reference sequence

16. Select the 5 entries in the *Database entries* panel that you want to include in the SNP analysis. Make sure the reference sequence is not selected and select ***Analysis*** > ***Sequence read set types*** > ***Map to reference***.

The *Map to reference* dialog box will open, containing several tabs (see Figure 17).

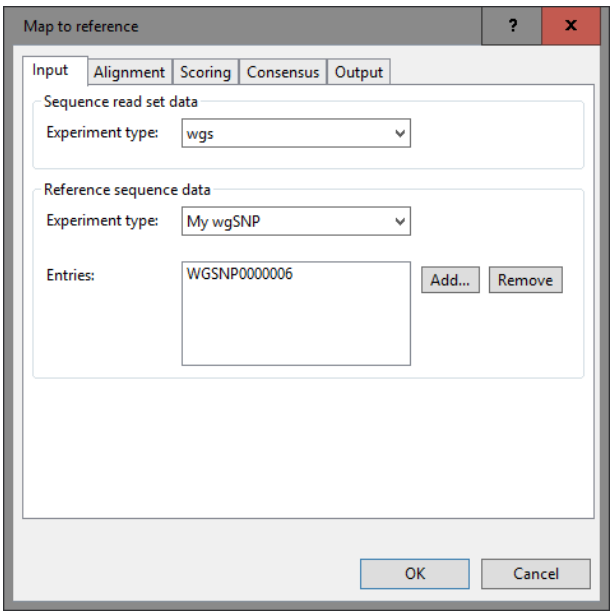


Figure 17: The *Map to reference* dialog box.

The map to reference action always works on the selected entries and on the sequence read set experiment type specified under ***Sequence read set data***. The reference sequence(s) to map against can be selected under ***Reference sequence data***. If a reference mapped sequence type is selected from the ***Experiment type*** drop-down list, the corresponding reference sequence will automatically be selected in the ***Entries*** list.

In the *Alignment tab*, settings for the alignment algorithm can be specified, the *Scoring tab* contains the settings to assess the initial alignment. The *Consensus tab* groups the settings for calculating a consensus sequence based on the final alignment.

In the *Output tab*, the experiments should be specified in which the consensus sequences should be stored.

17. Make sure the ***wgs*** is selected as experiment containing the read sets and the ***My wgSNP*** is selected as experiment containing the reference sequence. The entry holding the reference sequence is automatically displayed.

18. Press <***OK***> to start the mapping of the sequence read sets of the selected entries against the selected reference sequence.

The calculations might take several minutes. The resulting consensus sequences are stored in the appointed sequence type experiment, here: ***My wgSNP***.

19. Click on a green colored dot in the *Experiment presence* panel corresponding to the ***My wgSNP*** experiment of one of the selected entries to open the *Sequence editor* window containing the consensus sequence (see Figure 18).

20. The Power Assembly Project containing the Assembly can be called with ***File*** > ***Open assembler*** ( ). Alternatively a Power Assembly Project can be opened by double-clicking the project in the *Power assemblies* panel in the *Main* window.

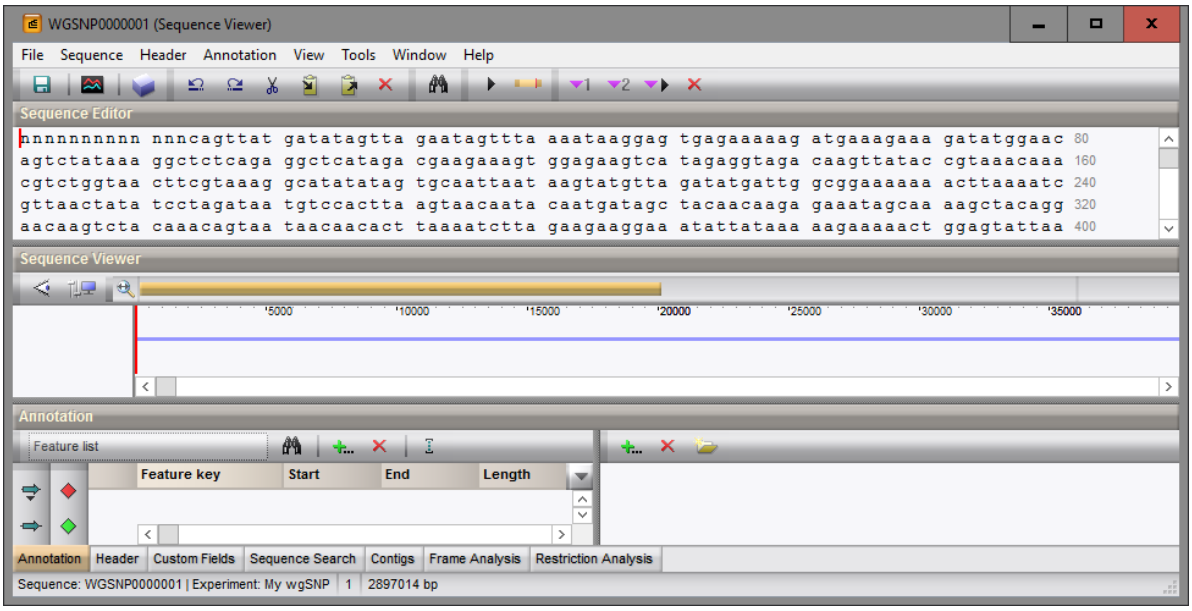


Figure 18: The Sequence editor.

Since Power Assembly Projects quickly fill up the database, it is recommended to delete these Projects when they become obsolete. Deleting Power Assembly Projects can be done in the *Power assemblies* panel (***Edit*** > ***Delete selected objects...*** ()). This action will only remove the assembly and not the resulting consensus sequence.

21. Close the Power Assembly Project and the *Sequence editor* window.

3.4 Perform wgSNP analysis and filter out revelant SNPs

22. Select the 5 entries in the *Database entries* panel that you want to include in the SNP analysis and select ***Analysis*** > ***Sequence types*** > ***Start SNP analysis...*** to start the *SNP analysis* wizard.

23. Select ***My wgSNP*** as ***Experiment type*** and press <***Next***>.

A number of predefined SNP templates are available.

24. Highlight the ***Strict filtering*** template and press <***Next***> (see Figure 19).

25. Check ***Open SNP analysis window*** and press <***Finish***>.

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 20.

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.

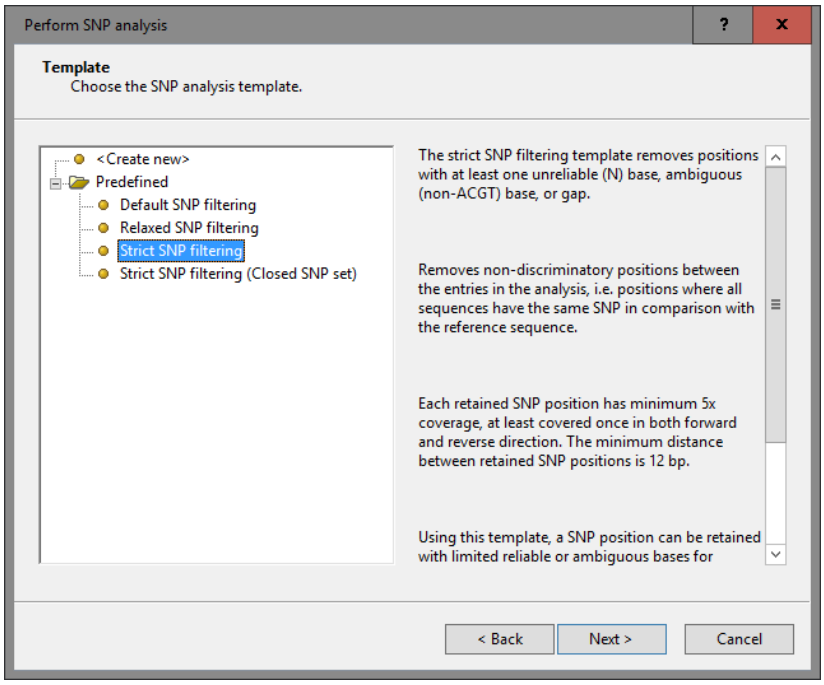


Figure 19: Choose a SNP filtering.

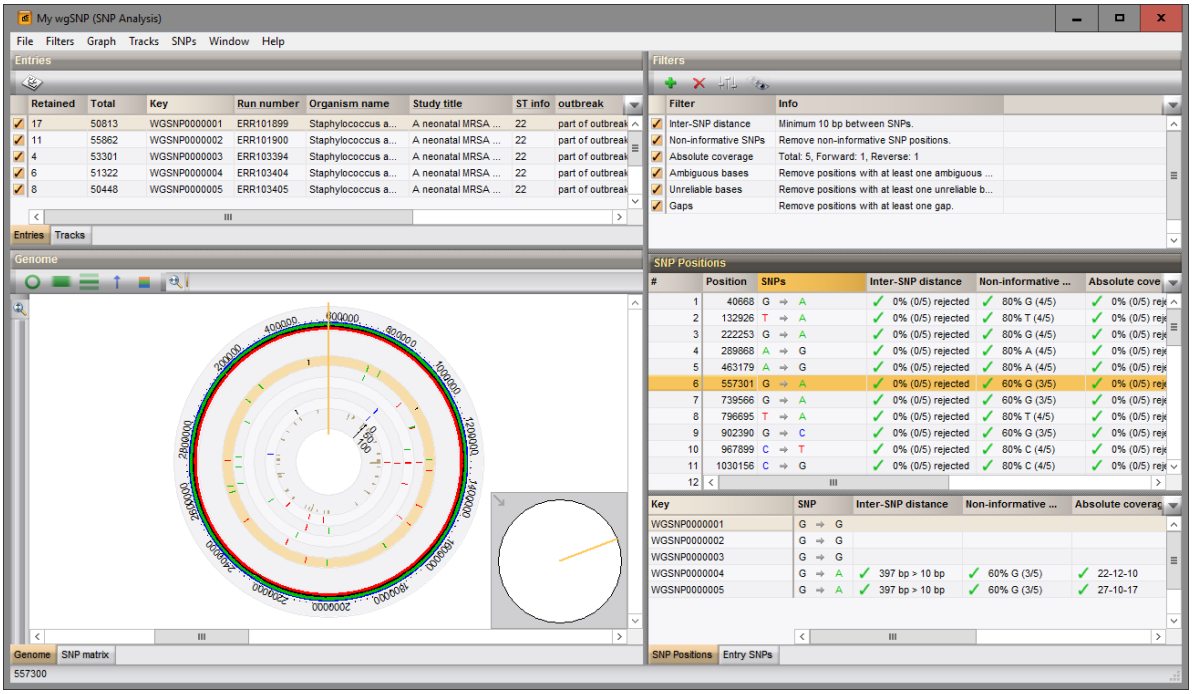


Figure 20: SNP analysis window.

- 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 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.
- 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:

26. 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.

27. 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.

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

29. 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.

30. Click on the tab of the *SNP matrix* panel to show the SNP matrix (see Figure 21).

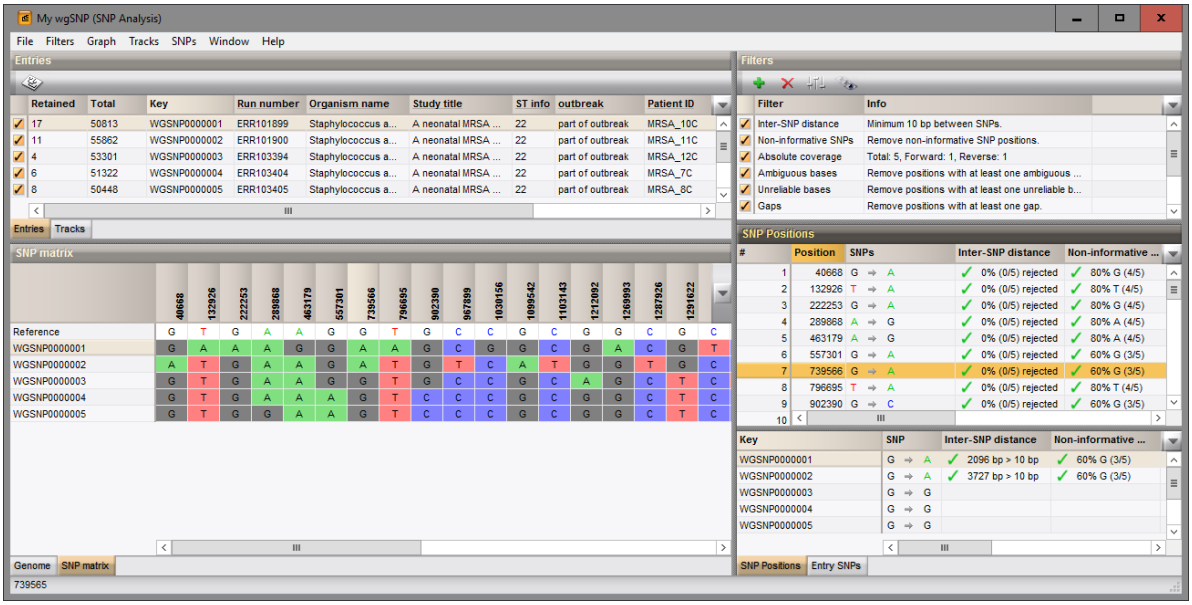


Figure 21: SNP matrix displayed.

31. 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.

4 Follow-up analysis

4.1 Cluster analysis on SNP data

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

In the comparison, the SNP matrix is available as a *character aspect* of the ***My wgSNP*** sequence experiment type (see Figure 22).

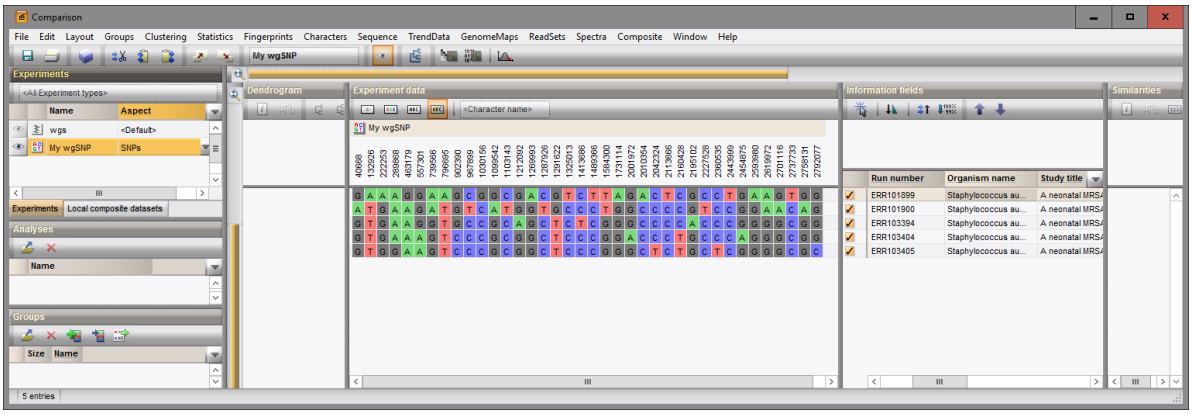


Figure 22: SNPs character aspect in the *Comparison* window.

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 **My wgSNP** 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 (differences)*** coefficient (see Figure 23).

The ***Categorical (differences)*** coefficient treats each different value as a different state, and results in a distance matrix.

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

The resulting dendrogram is displayed in the *Dendrogram* panel of the *Comparison* window (see Figure 24).

5. To view the number of SNPs on the branches, select ***Clustering*** > ***Dendrogram display settings...*** (), and tick the option ***Show node information***.

To trace back the number of SNPs from the branches or distance matrix, the displayed values needs to be multiplied with the ***Scaling factor*** used (default value = 1).

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

4.2 Exporting SNP data

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

7. In the *Comparison* window, select **SNPs** from the 'Aspect' drop-down list next to ***My wgSNP***.

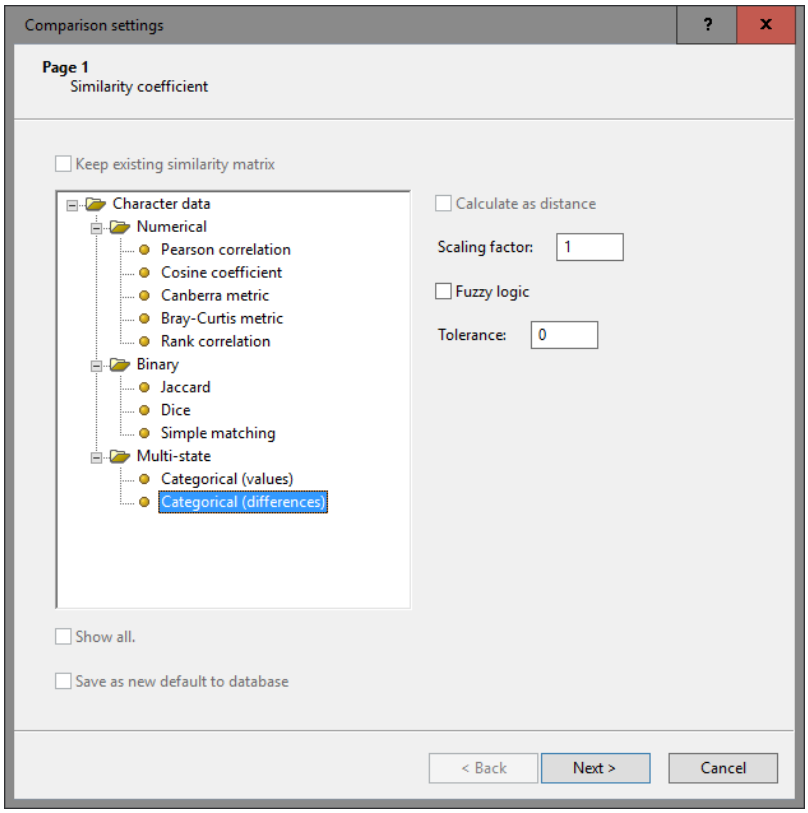


Figure 23: Similarity coefficient.

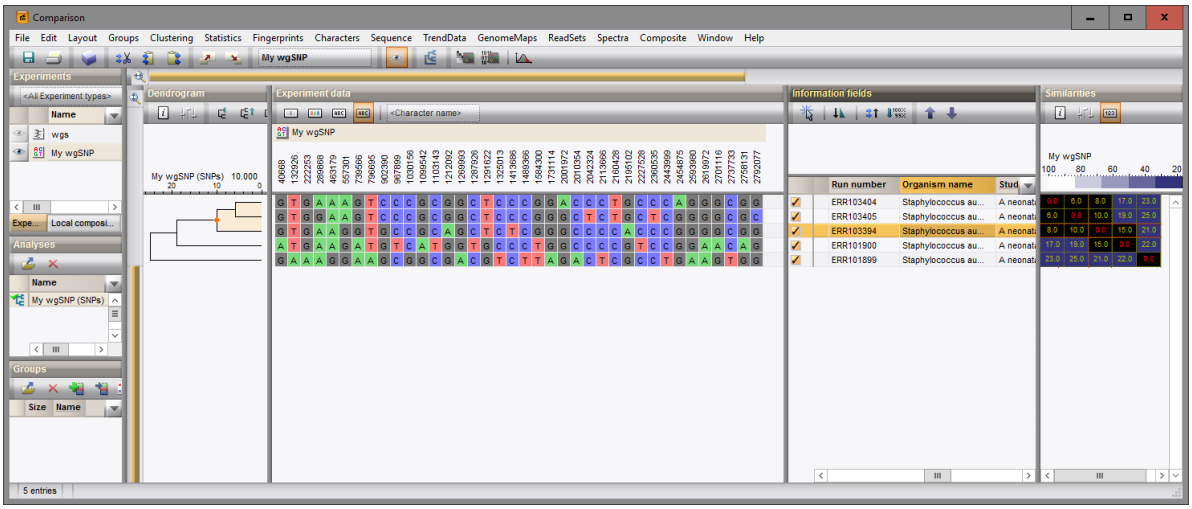


Figure 24: Dendrogram based on the SNP matrix.

8. Select ***File*** > ***Export*** > ***Export character data...***.

9. 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:

10. In the *Comparison* window, with the **SNPs** 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.