BioNumerics Tutorial:

Sequence typing of polymorphic VNTRs

1 Aim

Variable-Number Tandem Repeats (VNTRs) are well known for their high mutation rate and are therefore widely used for subtyping. Some VNTRs, however, exhibit polymorphism in their individual repeat sequences. By sequencing the polymorphic repeat region, each new repeat variant determined can be assigned a unique repeat code. The repeat succession for a given strain in turn, determines that strain's VNTR type.

In this tutorial you will learn how to install and use the *Polymorphic VNTR typing plugin* to explore polymorphic VNTR regions in some imported sequences.

The best known example is undoubtedly *spa-typing*, which is widely used for sub-typing of *Staphylococcus aureus*. Because of the specificity of spa-typing in terms of standardization, server synchronization and various other settings, a separate *spa typing plugin* is available in BioNumerics.

2 Example data

In this tutorial, an example data set is used which can be found on the Applied Maths website (http://www.applied-maths.com/download/sample-data, click on "TRST sample data files"). The data set was obtained from the Dublin Dental School and Hospital, Dublin, Ireland.

The sequence files originate from *Staphylococcus aureus* and represent the mec-associated Direct Repeat Unit (DRU) region.

3 Preparing the database

1. Create a new database (see tutorial "Creating a new database") or open an existing database.
2. Select ***File*** > ***Install / remove plugins...*** ().
3. Select the *Polymorphic VNTR typing plugin* from the list in the *Applications tab* and press the <***Activate***> button.
4. The program will ask to confirm the installation of the plugin. Press <***OK***> twice to confirm the installation.
5. When the *Polymorphic VNTR typing plugin* is successfully installed, a confirmation message pops up. Press <***OK***>.
6. Press <***Proceed***> (or <***Exit***>) to close the *Plugins* dialog box and to continue to the *Main* window.
7. Close and reopen the database to activate the features of the *Polymorphic VNTR typing plugin*.

The *Polymorphic VNTR typing plugin* installs menu items in the main menu of the software under ***Repeat-Typing***.

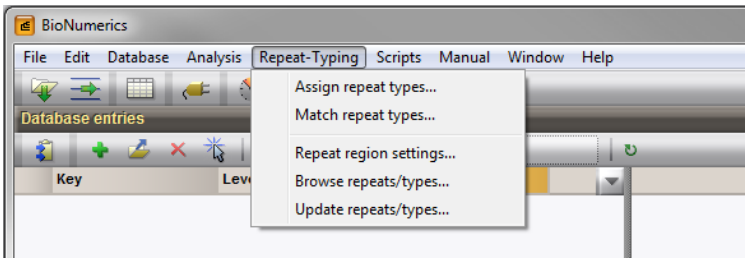

Figure 1: Repeat menu items in the *Main* window.

4 Repeat type settings

1. Select ***Repeat-Typing*** > ***Repeat region settings*** to call the *Repeat regions* dialog box.
2. Select the <***Add new***> button in the *Repeat region panel*.
3. For this exercise enter the name ***Dru*** and press <***OK***>.
4. Optionally enter a ***Description*** in the *Repeat region panel*.
5. Enter the start pattern `GATTATACTA` and stop pattern `ATAAGGGGTACAGAAAAACT` in the *Experiments panel*.
6. For this exercise, enter the following URLs in the *Update repeats/types panel*:
   - ***Repeats***: `http://www.dru-typing.org/downloads/drurepeats.txt`
   - ***Types***: `http://www.dru-typing.org/downloads/drutypes.txt`
7. Leave all the settings unaltered in the other panels (see Figure 2) and press <***OK***>.

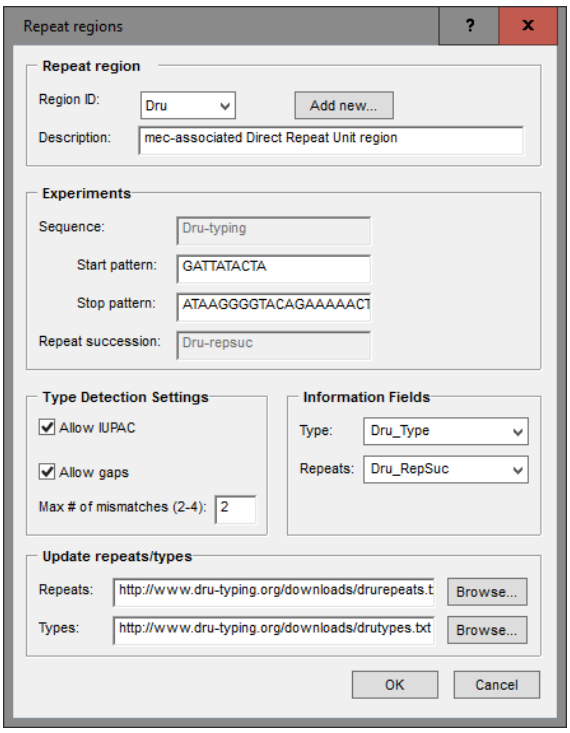

Figure 2: Repeat settings.

The information fields specified in the *Repeat region dialog box* are created and are displayed in the *Database entries* panel of the *Main* window.

BioNumerics automatically creates a sequence type for the import and storage of sequence data (**Dru-typing**), and a character type for the storage of the repeats (**Dru-repsuc**). The experiments are listed in the *Experiment types* panel (see Figure 3).

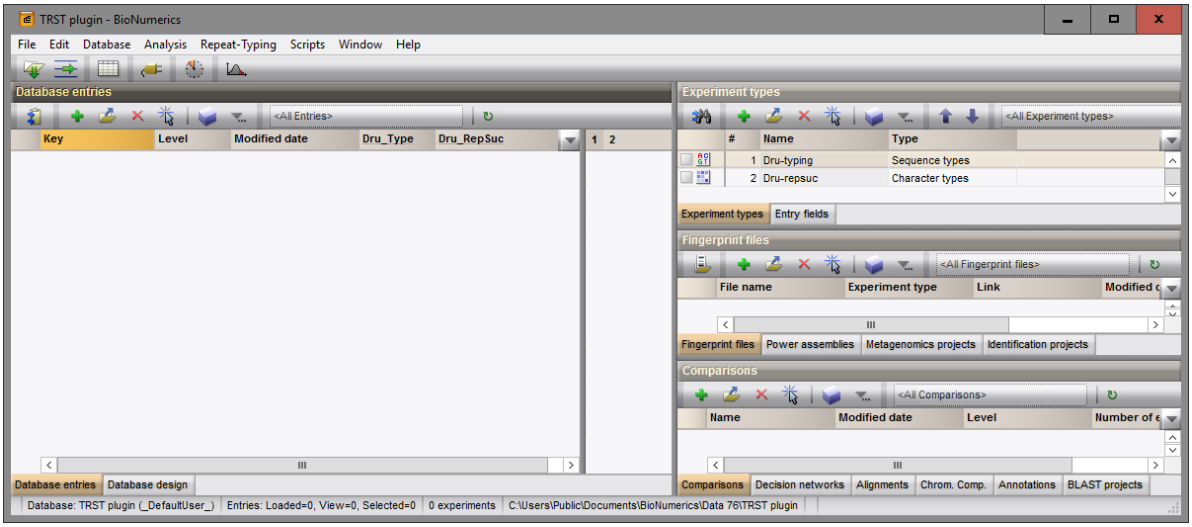

Figure 3: The *Main* window after installation of the plugin.

Repeat and type information available at the Dru Server can be uploaded to the BioNumerics database.

8. Select ***Repeat-Typing*** > ***Update repeats/types*** to update the repeats and/or types.

The repeat and type lists are updated. A confirmation message pops up.

9. Press <***OK***> once more.

The uploaded repeats and repeat types can be queried from within BioNumerics.

10. Select ***Repeat-Typing*** > ***Browse repeats/types*** in the *Main* window.

This action calls the *Browse types/repeats* dialog box (see Figure 4).

11. Click on ***types*** to view the list of repeat types.

12. Close the *Browse types/repeats* dialog box.

It is also possible to view all repeats and types stored in the database with an *object query*.

13. In the *Main* window, select ***Database*** > ***Object queries...*** ( ) and select "<Create new>" from the drop-down menu that appears. Press <***OK***>.

14. As ***Object to report***, select "TRST:Repeat sequences" or "TRST: Sequence types" and press <***OK***> (see Figure 5 for an example).

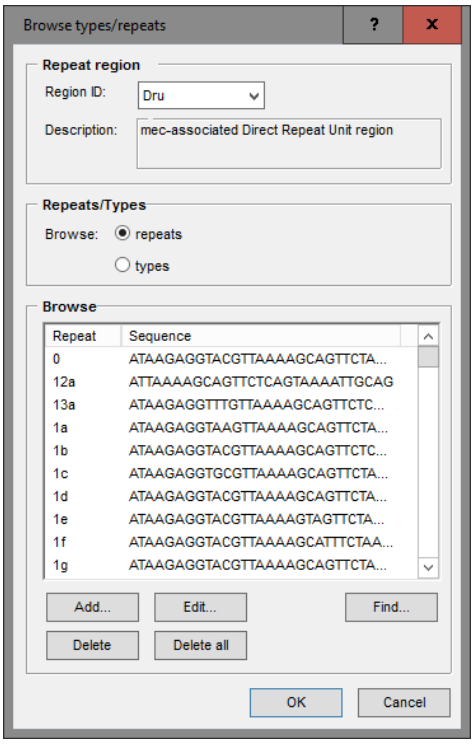

Figure 4: Browse repeats and types.

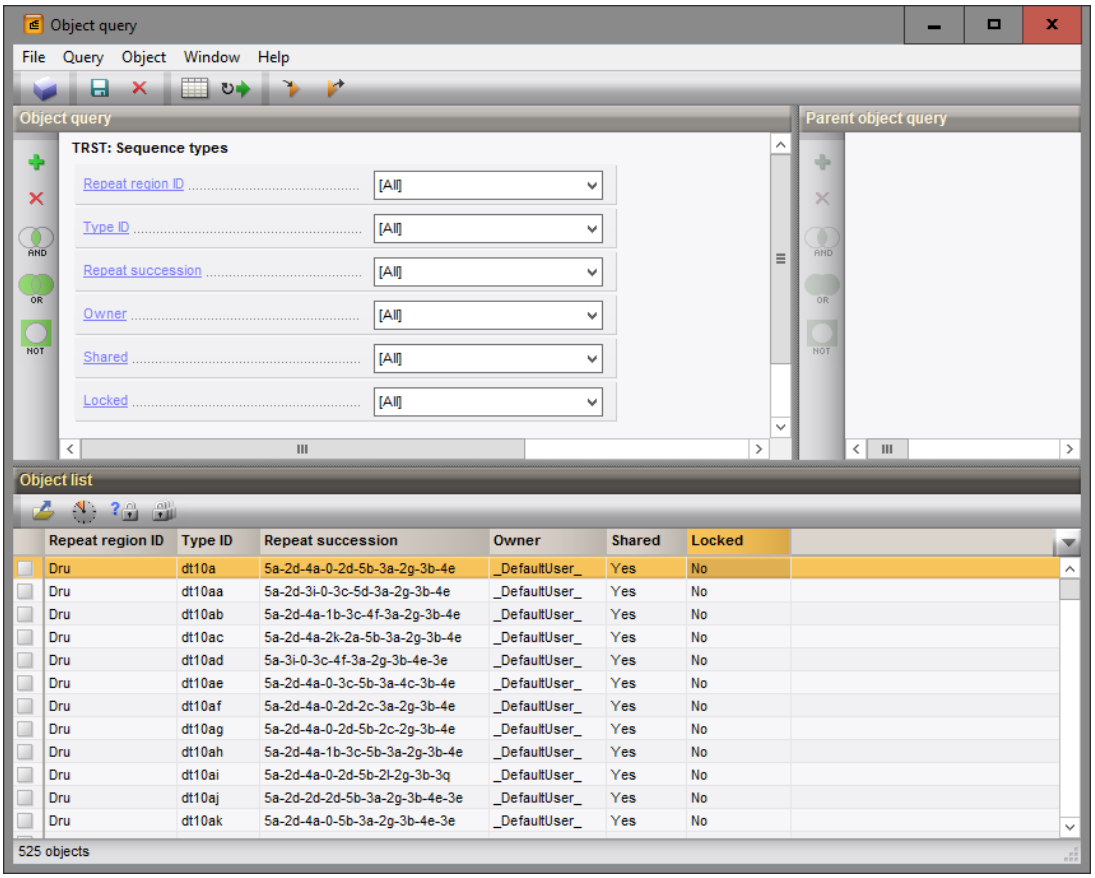

Figure 5: Object query of the sequence types.

5 Importing and assembling trace files

5.1 Importing and assembling trace files in batch

A set of trace files can be downloaded from the Applied Maths website (http://www.applied-maths.com/download/sample-data, click on "TRST sample data files") and are used in this tutorial.

1. Select ***File*** > ***Import...*** (, **Ctrl+I**) to call the *Import* dialog box.

2. Select ***Import and assemble trace files*** under ***Sequence type data*** and press <***Import***>.

3. Select the <***Browse***> button, navigate to the correct path, select all the sequence trace files and press <***Open***>.

The *Import sequence traces* wizard page is updated (see Figure 6).

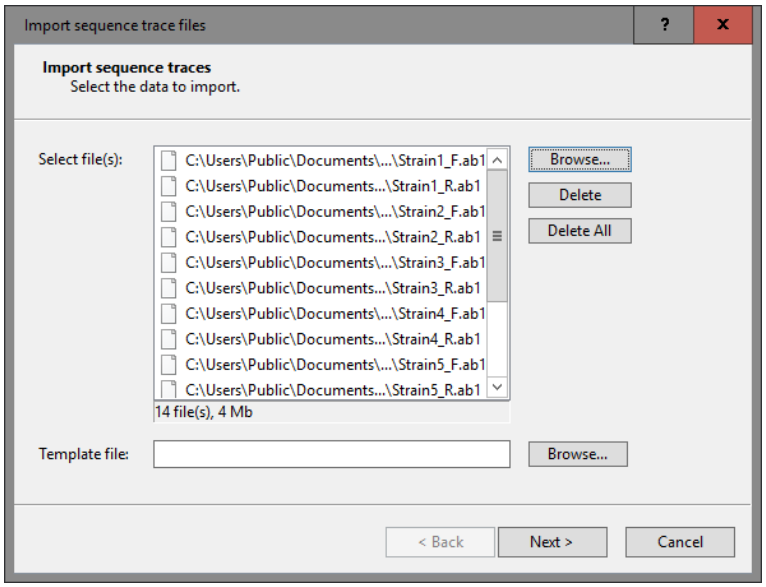

Figure 6: Select trace files.

4. Press <***Next***> to go the next step.

The way the information should be imported in the database can be specified with an import template. In the example data set, the ***Key*** is provided in the trace file name. The import template ***Example import 1*** parses the strain name from the file names and saves it to the ***Key field***.

5. Select the ***Example import 1*** template and press the ***Preview*** button to have a look at the parsed information.

6. Close the preview.

7. Select the ***Dru-typing*** experiment from the ***Experiment type*** list and press <***Next***> (see Figure 7).

8. Press <***Next***>.

9. Press <***Next***> once more to confirm the creation of 8 new entries (see Figure 8).

The *Processing* wizard page opens (see Figure 9).

In the *Reports panel*, the ***Maximum# of unresolved bases reported*** can be specified (default value 20). Likewise, the ***Maximum # of align inconsistencies reported*** can be entered (default value 20). Align

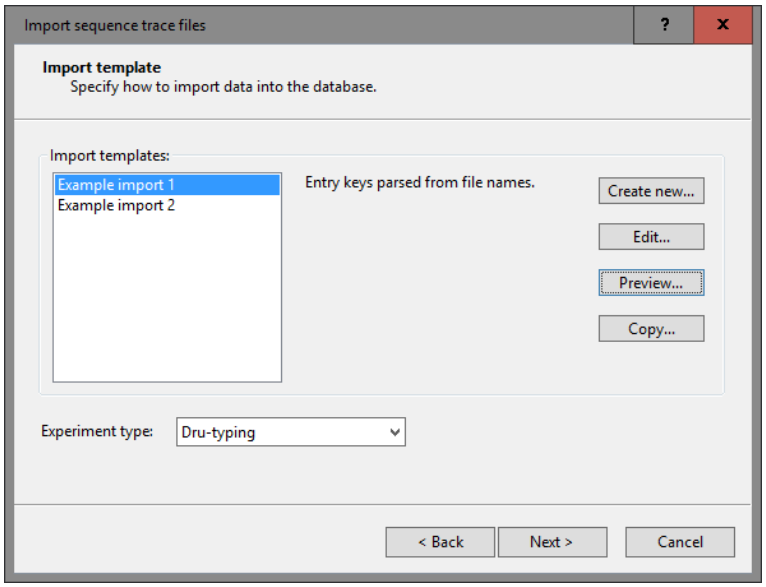

Figure 7: Import template.

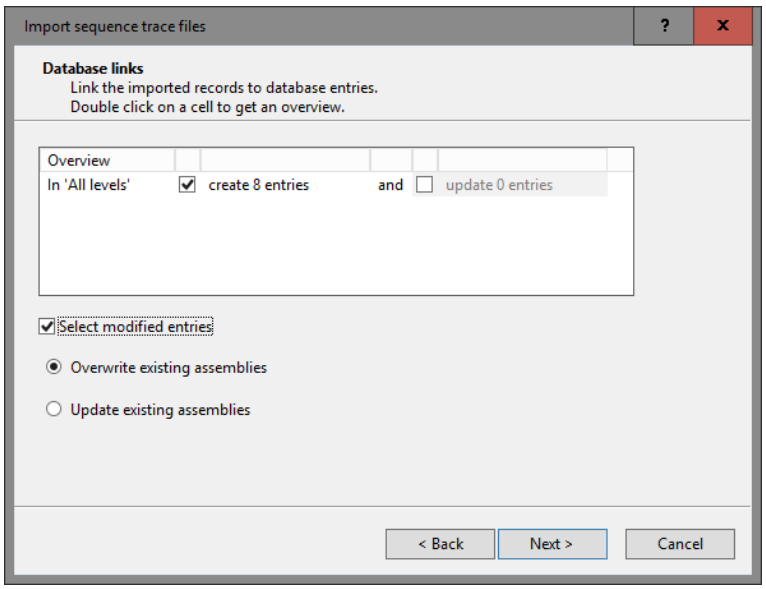

Figure 8: Database links.

inconsistencies are positions where the consensus is resolved, but where one or more sequences are different from the consensus.

10. Press <***Trimming settings***> to pop up the *Assembly trimming settings* dialog box (see Figure 10).

Following settings can be specified:

- ***Minimum # of sequences*** specifies the minimum number of trace sequences that should contribute to the subsequence on the consensus that matches the trimming targets. For example, if "2" is entered, a trimming target will only be set if the matching region on the consensus is *fully* defined by at least 2 sequences.
- For both the ***Start position*** and ***Stop position***, a ***Trim pattern*** is displayed. The use of IUPAC code for ambiguous positions is supported. The ***Tolerance*** defines the number of mismatches allowed for

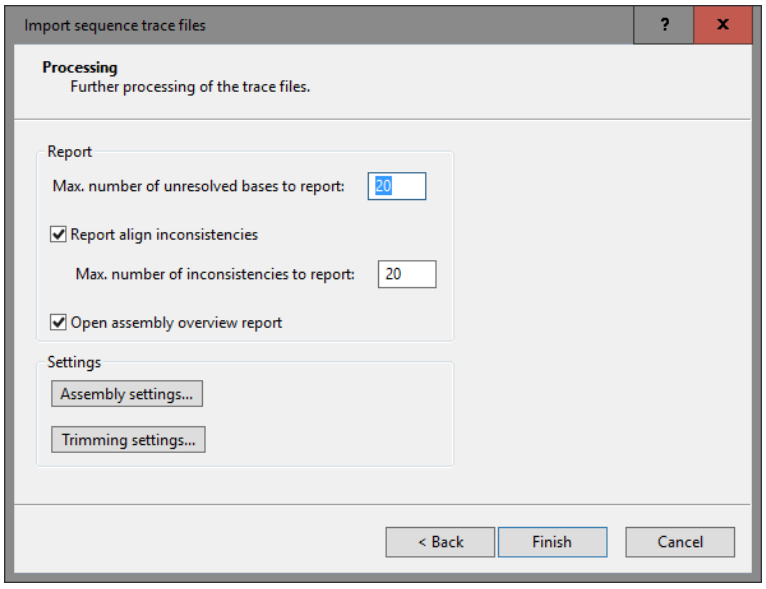

Figure 9: The *Processing* wizard page.

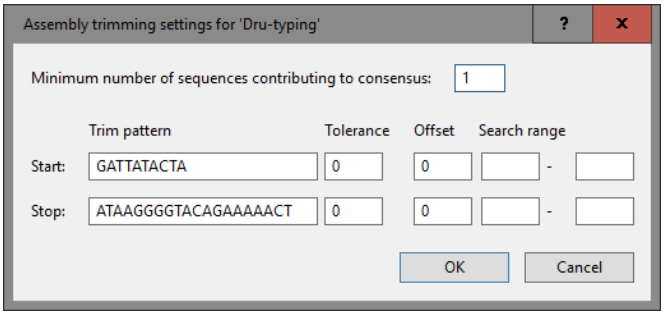

Figure 10: Trimming patterns.

a sequence to be recognized as a trim pattern. With the ***Offset***, one can specify that the consensus is trimmed at a certain offset from the start and end trimming target positions. If no offset is specified (zero), the trimming targets are included in the trimmed consensus. With the ***Search range*** one can restrict the search to certain regions on the consensus, e.g. to prevent incidental matches inside the targeted consensus sequence.

The entered trim patterns will be searched on the consensus sequence in both directions, i.e. on the consensus as it appears as well as on its complementary strand. In case the trim patterns match the complementary strand of the consensus, it will be automatically invert-complemented. If the ***Trim pattern*** text boxes are left empty, no preference sense is available.

The trimming patterns entered in the *Repeat regions* dialog box for the sequence type ***Dru-typing*** (see 4) are shown in the ***Start pattern*** and ***Stop pattern*** text boxes.

11. Leave the predefined settings unaltered and press <***OK***> to close the trimming dialog box.

12. Press the <***Assembly settings***> button to call the *Assembly settings* dialog box (see Figure 11).

The Assembly settings are grouped in tabs per settings dialog box in *Assembler*: ***Quality*** assignment, ***Assembly*** and ***Consensus*** determination. For a detailed description of the Assembler program settings, see the reference manual. In the last tab the Assembly settings can be copied from or to another sequence type

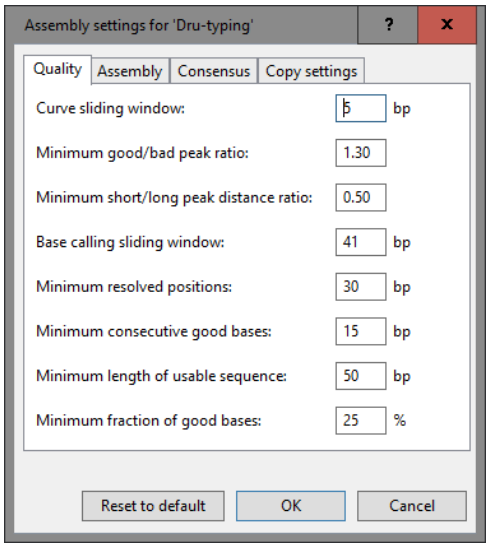

Figure 11: Assembly settings.

experiment.

13. For this exercise, do not change the settings and press <***OK***>.

14. Make sure the option ***Open assembly overview report*** is checked and press <***Finish***> to assemble the selected trace files from the example dataset into separate contig projects.

5.2 Reports

When the assemblies are processed, an interactive report window appears (see Figure 12). This window can also be displayed from the *Main* window with ***Analysis*** > ***Sequence types*** > ***Batch assembly reports...***.

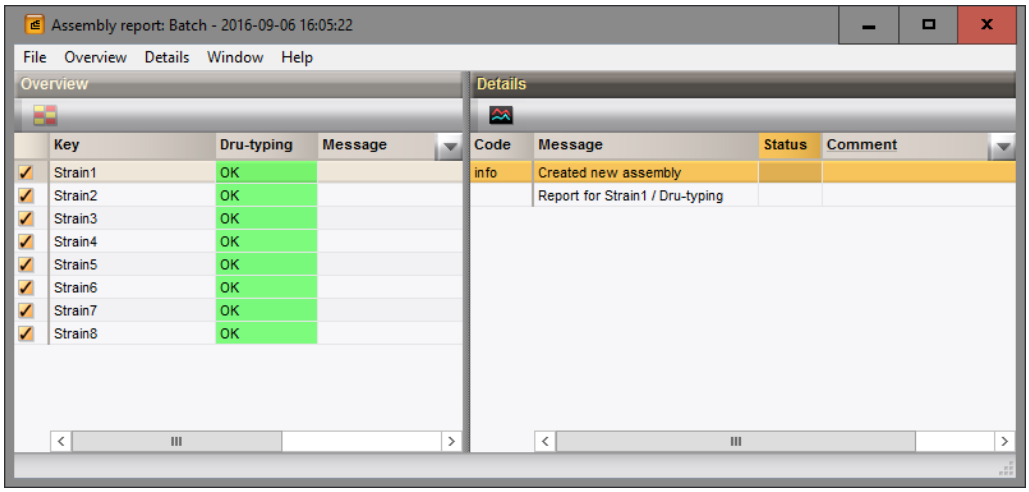

Figure 12: Overview report.

The *Overview panel* displays the entries (keys) as rows and the experiments as columns. Each cell, corresponding to a key/experiment pair, provides information about the current status of the contig project. This information can be:

- **N/A**: No such experiment exists with this key.

- **N/B**: An experiment with this key exists, but (a) the assembly was not created from this batch; or (b) no batch sequence assembly is present for this sequence.
- **OK** (green): A contig was assembled without any problems.
- **Warning** (orange): Align inconsistencies occurred that were resolved under the applied consensus determination settings.
- **Error** (red): At least one of several possible assembly errors occurred, e.g. a trace sequence did not meet the quality criteria, more than one contig was created, the trimming positions were not found or unresolved bases are present in the consensus.
- **Solved** (green): A warning or error that was solved by the user.

15. Click a cell, e.g. ***Strain1/Dru-typing*** to update the *Details* panel on the right-hand side.

The *Details* panel is organized in message rows with four columns.

- The first column displays a message **Code**, which can be either ”info”, ”warning” or ”error”.
- The second column shows the actual **Message**. Double-clicking on this cell opens the *Contig assembly* window (if not already open), with the corresponding position highlighted.
- The third column displays the **Status** of the message, which can be ”new”, ”read” or ”solved”. The status can be changed by the user.
- The fourth column is a **Comment** field. A comment can be entered by the user.

6 Checking assemblies in Assembler

6.1 Introduction

The *Contig assembly* window can be launched from the *Batch sequence assembly report* window or from the *Main* window:

- Double-click on a message cell in the *Details* panel of the *Batch sequence assembly report* window of an key/experiment combination to launch Assembler.
- As soon as an experiment is linked to a database entry, the *Experiment presence* panel shows a colored dot for the experiment of this entry. Click on the colored dot in the *Experiment presence* panel while holding the **Shift**-key to open the *Experiment card* window for an entry. In the *Experiment card* window, click on the  button to launch Assembler.

1. Open the *Contig assembly* window for the entry with key **Strain1** by double-clicking on the first message in the *Details* panel of the *Batch sequence assembly report* window.

The *Alignment* panel in the *Contig assembly* window shows the consensus sequence (upper line) and the individual trace sequences that contribute to the displayed consensus. The upper panel (*Alignment overview* panel) displays the aligned trace sequences. If the arrow points to the left, the program has invert-complemented the sequence to obtain the correct alignment. The upper left panel displays the selected consensus with its length and the number of sequences that are part of it.

2. Select the *Aligned traces* panel.

The bottom panel now displays the chromatogram files for both trace sequences (see Figure 13).

3. To obtain an optimal view of the curves, use the zoom sliders in the *Traces* panel or use the zoom buttons.

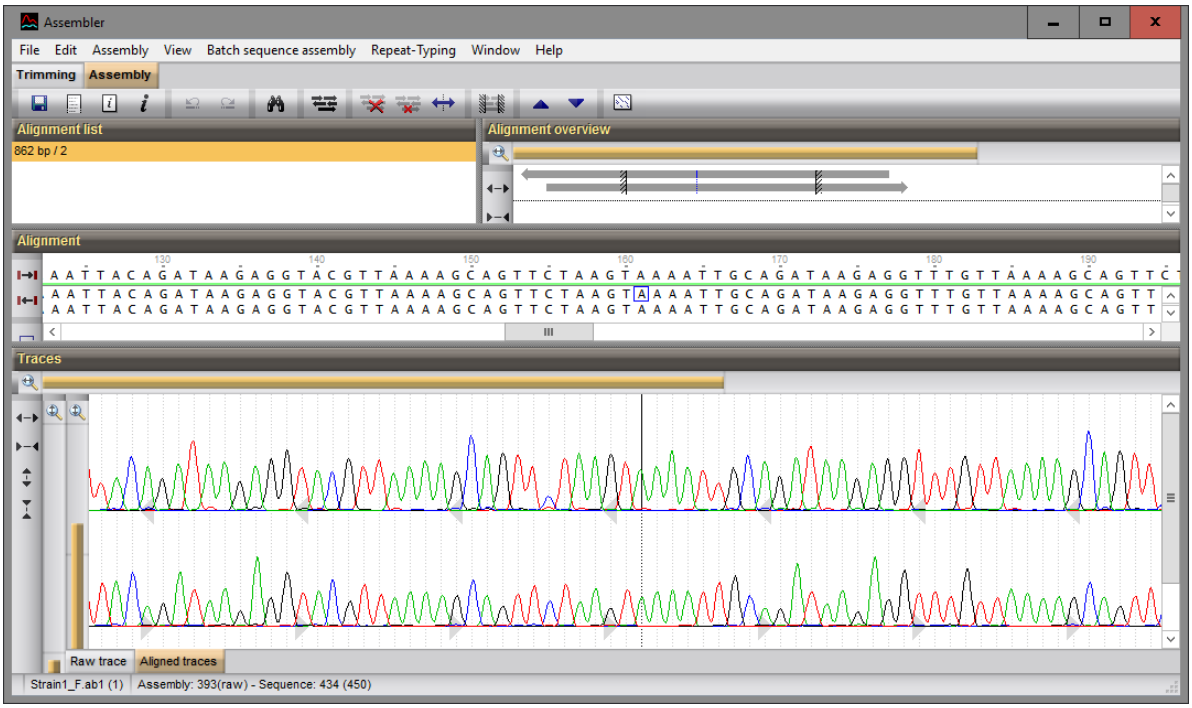

Figure 13: The *Aligned traces* panel.

6.2 Showing repeats on the consensus

4. In the *Contig assembly* window, select ***Repeat-Typing*** > ***Show repeats*** or use the shortcut **Shift+F5**.

Assembler screens the consensus sequence for repeats.

- Known repeats are shown in *green* and the name of the repeat is shown on top of the know repeat sequence.
- Bases in the repeat succession string that are not assigned to a known repeat are shown in red.
- The 5' and 3' signatures are displayed in *yellow*.

If the option ***Allow IUPAC code*** is checked in the *Repeat regions* dialog box and *one of the bases* of a IUPAC code in the consensus results in a match with a known repeat, the repeat is shown in green and the name of the repeat is shown on top of the corresponding repeat sequence in the *Alignment* panel.

The repeat succession string and the corresponding repeat type (if present in the repeat type list) are displayed in the caption of the *Contig assembly* window.

When importing and assembling sequences, BioNumerics uses the parameters defined in the *Assembly settings* dialog box.

5. Select ***File*** > ***Show report*** ( ) to view all parameters.

After import, these parameters can still be changed for each individual assembly.

1. Select the *Trimming* panel and select ***File*** > ***Quality assignment...*** ( ) to change the quality assignment settings. This action can only be used if the alignment is removed.
2. Select the *Assembly* panel and choose ***Assembly*** > ***Assemble sequences...*** ( ) to change the assembly settings.

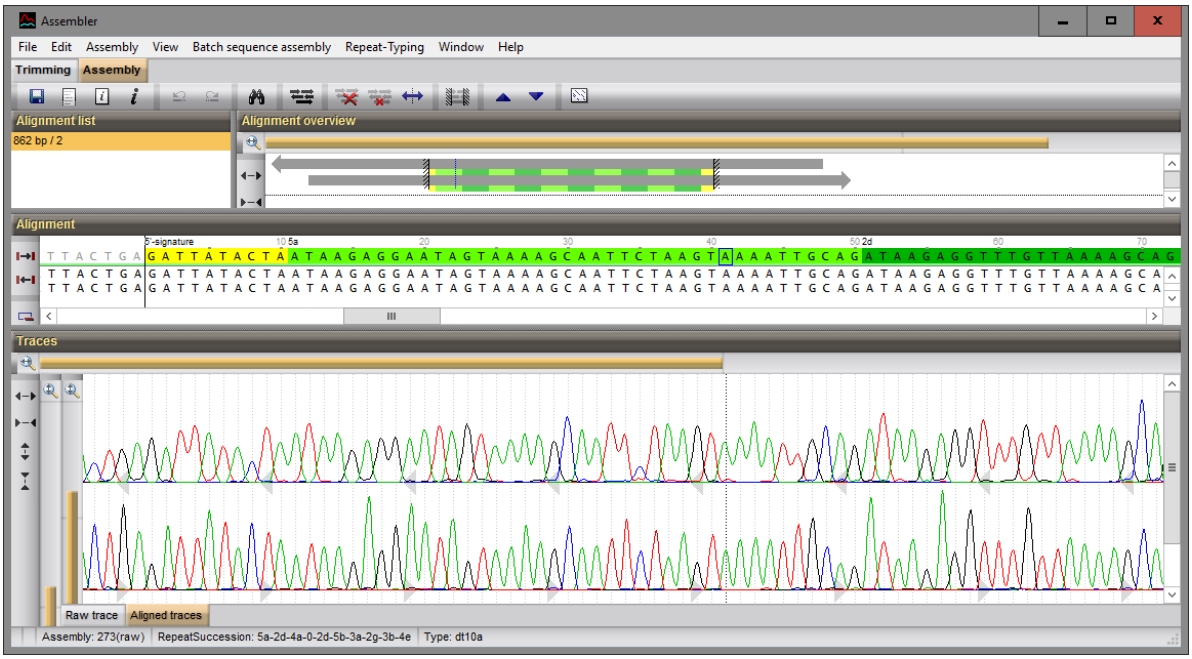

Figure 14: Showing the repeats on the consensus sequence.

3. If you want to change the Consensus determination parameters, select the *Assembly* panel and select ***Assembly*** > ***Consensus determination...***.

Detailed information on each of these parameters can be found in the reference manual.

6.3 Showing the repeat succession plot

6. Select ***Repeat-Typing*** > ***Show repeats plot*** or use the shortcut **Shift+F6**.

The repeats are displayed in the *Repeat plot window* (see Figure 15).

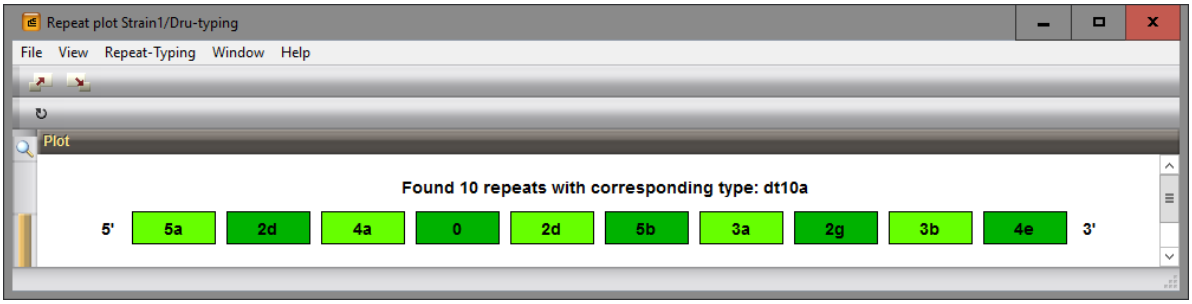

Figure 15: Repeat plot.

When clicking on an unknown red "r??" repeat, a table is displayed with suggestions to edit the sequence (see Figure 16). In the left column, the repeat is shown. In the right column, the associated repeat type - if available - is displayed.

7. Use the zoom functions and (***View*** > ***Zoom in*** and ***View*** > ***Zoom out***) to obtain the best view of the plot.

Replacing the "T" with an "A" in the unknown repeat in Figure 16 results in repeat **5a** and repeat type **dt10a**.

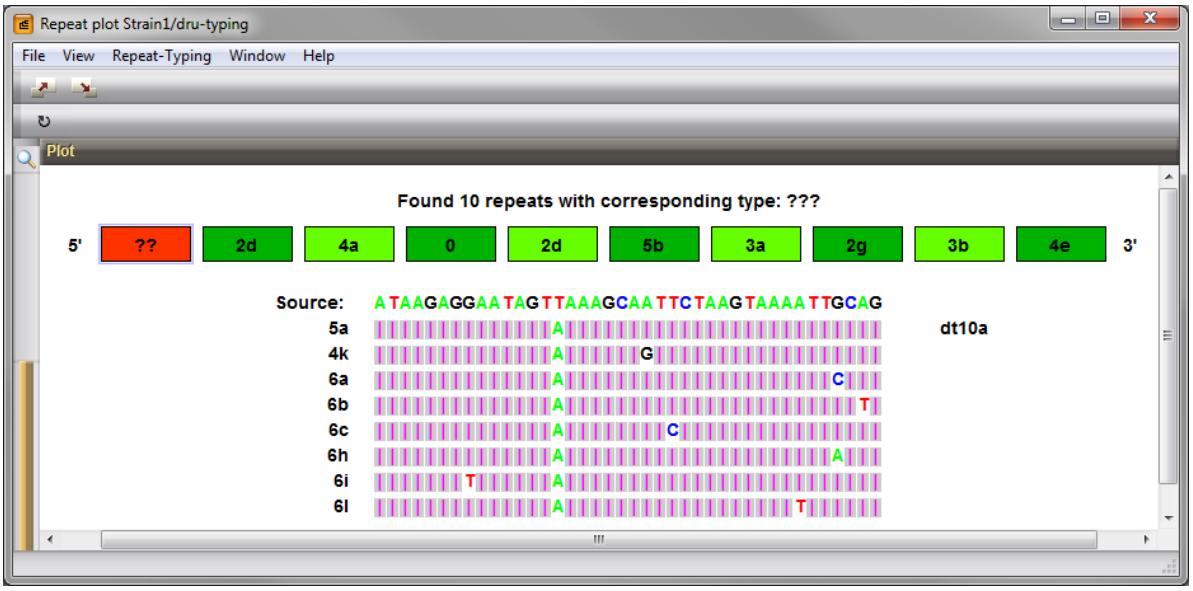

Figure 16: The repeat plot: editing suggestions are displayed for the unknown repeat.

Looking at this position in the *Assembly view* gives additional information about the missing base.

8. When the consensus sequence has been edited in the *Contig assembly* window, select ***Repeat-Typing*** > ***Refresh*** in the repeat plot to update the repeat information.

More information on how to edit sequences in Assembler can be found in the reference manual.

9. To copy the repeat plot to the clipboard, select ***File*** > ***Copy to clipboard***.

10. The plot can be printed with ***File*** > ***Print***.

11. Close the *Repeat plot window* with ***File*** > ***Exit***.

6.4 Changing the status of warning and error messages

Only for those entries that have a green (= **OK** or **Solved**) or orange (= **Warning**) status, the repeat types can be assigned.

- It is recommended to check the *warning* messages and solve them if needed. Since repeat types can be assigned to entries that have a Warning status, it is not required to change the status to ”Solved”.
- *Errors* need to be checked in the *Contig assembly* window and solved. Since repeat types cannot be assigned to entries that have an Error status, it is required to change the status to ”Solved” after having solved all errors in Assembler.

12. Select ***Batch sequence assembly*** > ***Set report to solved, save and close*** (**Ctrl+Shift+S**) in the *Contig assembly* window.

The corresponding key/experiment cell in the overview *Batch sequence assembly report* window is updated and displayed in green. The status ”Solved” is displayed in the key/experiment field.

13. After having solved all warnings and/or errors in Assembler, select ***File*** > ***Save*** ( , **Ctrl+S**) and ***File*** > ***Exit*** to close the *Contig assembly* window.

14. In the *Batch sequence assembly report* window, select ***Details*** > ***Set message to solved*** (**S**).

The corresponding key/experiment cell in the *Overview* panel is updated and displayed in green. The status "solved" is displayed in the cell and in the ***Status*** column of the *Details* panel (see Figure 17 for an example).

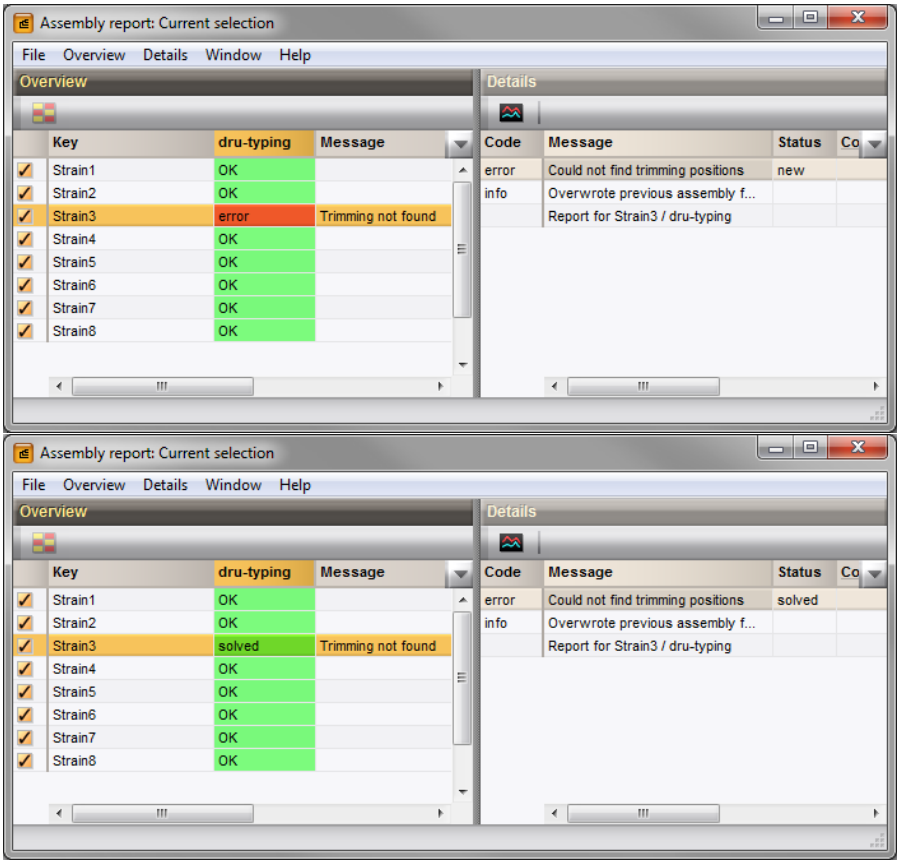

Figure 17: Solve errors/warnings.

7 Repeat typing in BioNumerics

7.1 Selections in the main window

In the *Main* window, a repeat typing experiment (in our example: **Dru-typing**) is present for each of the assembled sequences (see colored dot in the second column in the *Experiment presence* panel).

Screening for repeats and types can be done for all entries present in the database, or for any selection of entries in database.

1. Select a single entry in the *Database entries* panel by holding the **Ctrl**-key and left-clicking on the entry. Alternatively, use the **space bar** to select a highlighted entry or click the ballot box next to the entry.

Selected entries are marked by a checked ballot box (☑) and can be unselected in the same way.

2. In order to select a group of entries, hold the **Shift**-key and click on another entry.

A group of entries can be unselected the same way.

3. All entries can be selected at once with ***Edit*** > ***Select all*** (**Ctrl+A**).

4. Clear all selected entries with ***Database*** > ***Entries*** > ***Unselect all entries (all levels)*** (, **F4**).

7.2 Assigning types

5. Make a selection in the *Main* window. To select all entries at once, use ***Edit*** > ***Select all*** (**Ctrl+A**).

6. Select ***Repeat-Typing*** > ***Assign repeat types*** in the *Main* window and confirm the assignment.

If no selection is present in the database, the software will display a message asking you if you wish to run the tool on the complete database.

The *Polymorphic VNTR typing plugin* uses a 2-step approach when the command ***Repeat-Typing*** > ***Assign types*** is selected:

7.2.1 Step 1: The assembly is screened for repeats

The repeat succession is stored in the character type ”repeatID-repsuc” (in this exercise: ***Dru-repsuc***) *and* the succession is displayed in the database information field that holds the repeat succession information (in this exercise: ***Dru_RepSuc***) (see Figure 18).

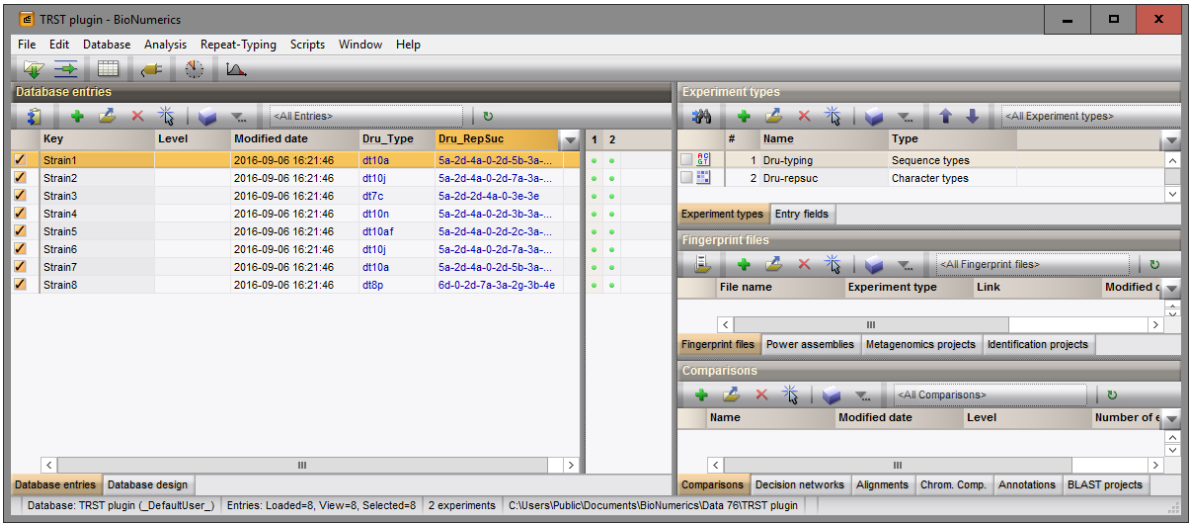

Figure 18: The *Main* window after repeat typing.

7. Click on the colored dot in the ***Dru_RepSuc*** column of the *Experiment presence* panel to open the character *Experiment card* window for an entry (see Figure 19).

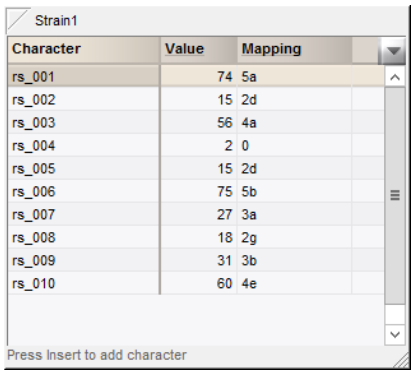

Figure 19: The character card.

8. Close the experiment card by clicking in the small triangle-shaped button in the left upper corner.

When a repeat does not match one of the repeats in the database, or when a IUPAC code is present in the consensus sequence, a ”??” is placed at this position in the repeat succession information field and in the ***Mapping*** column of the character card.

When a sequence is found that is too short or too long to be considered as a repeat sequence, an asterisk (*) is placed at this position in the repeat succession information field and in the ***Mapping*** column of the character card.

When no repeats are found, no information is written in the repeat succession information field.

7.2.2 Step 2: Repeat type (if available) is assigned to each selected entry

The repeat type is displayed in the information field that holds the repeats type information (in this exercise: **Dru_Type**).

The repeat type is denoted as ”???” if the repeat succession is incomplete. When the repeat information is currently not linked to a repeat type in the database, ”Unknown” is displayed in the repeat type information field. If no repeats are found, ”NA” (Not Available) is displayed.

8 Cluster analysis of repeat types

8.1 Introduction

In this chapter, we are going to take a look at the evolutionary relationship between the repeats by means of the construction of a dendrogram and a minimum spanning tree.

The *Polymorphic VNTR typing plugin* uses a multi-step approach for this cluster analysis.

- The plugin uses an algorithm based on a DSI model [1] for the pairwise alignment of the repeats. This *DSI model* considers three mutational events: Duplication of tandem repeats, Substitutions and Indels.
- Next, the cost matrix is used to correct for the evolutionary distances between the repeats.

Taking these costs into account, the output of the DSI model is a similarity matrix. From this similarity matrix, a dendrogram and/or a minimum spanning tree can be constructed.

8.2 The Comparison window

1. Make sure all entries are selected in the *Main* window (**Ctrl+A**).
2. Highlight the *Comparisons* panel in the *Main* window and select ***Edit*** > ***Create new object...*** () to create a new comparison for the selected entries.
3. Drag the separator lines between the panels to the left or to the right, in order to divide the space among the panels optimally.
4. Move the panels by clicking in the header of a panel and - while keeping the mouse button pressed - dragging it to another location in the *Comparison* window.

In our database, two experiment types are available and are shown in the *Experiments* panel.

5. Click on the eye button () of the character type ”regionID-repsuc” (in this exercise: **Dru-repsuc**).

The pattern images are displayed in the *Experiment data* panel. Initially, the character values are displayed as colors according to the color scale defined for each character (see the reference manual for more information).

6. Select ***Characters*** > ***Show mappings*** ( ) or ***Characters*** > ***Show mappings+colors*** ( ) to display the mapped name for each character value.

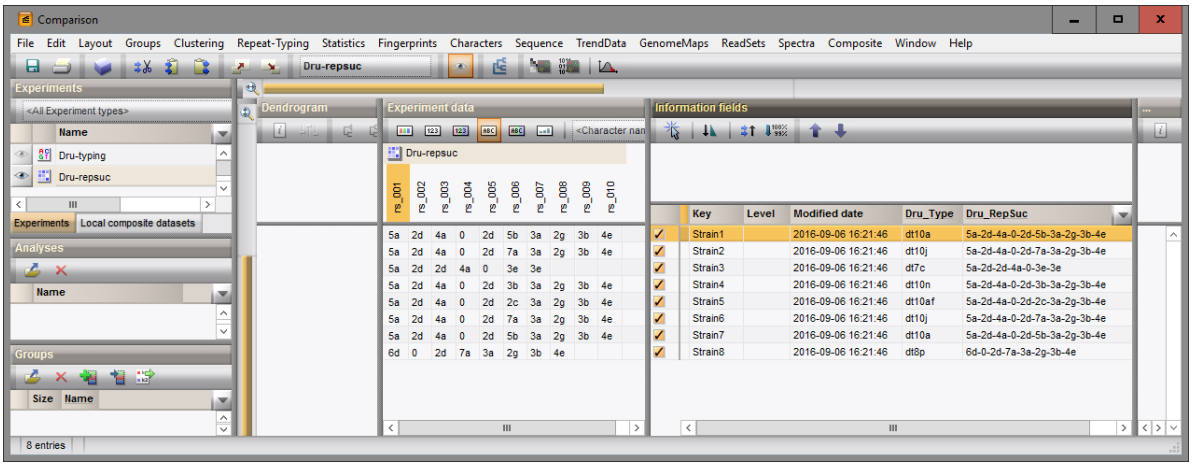

Figure 20: The *Comparison* window.

8.3 Creating a cost matrix

In the *Polymorphic VNTR typing plugin*, there is a default binary cost matrix available for the calculation of the dendrogram, consisting of two states: a match between the repeats and no match.

7. Select ***Repeat-Typing*** > ***Cost matrices*** in the *Comparison* window for the creation of your own cost matrix.

The *Cost matrices* dialog box displays all cost matrices defined by the user (initially empty).

8. Select <***Create new***>, specify a ***Cost matrix name***, specify the settings and press <***OK***>.

8.4 Cluster analysis

8.4.1 Settings

9. Select ***Repeat-Typing*** > ***Cluster types*** in the *Comparison* window.

The *Clustering* dialog box appears.

As an example, we will create a minimum spanning tree and a UPGMA dendrogram.

8.4.2 Minimum spanning tree

Minimum spanning trees are trees calculated from a distance matrix and possess the property of having a total branch length that is as small as possible. A MST chooses the sample with the highest number of related samples as the root node, and derives the other samples from this node. This may result in trees with star-like branches and allows for a correct classification of population systems that have a strong mutational or recombinational rate.

10. Select ***Repeat-Typing*** > ***Cluster types*** in the *Comparison* window.

11. Select ***Minimum Spanning Tree*** in the *Cluster Method panel*.

An additional setting called ***Distance bin size*** is displayed in the *MST panel*. Based on this setting, the software creates bins of certain distance intervals, that are converted into distance units. When for example the distance bin size is set to 1%, two entries having a similarity of 99.6% will have a distance of 0 (interval 100%-99% = distance 0). Two entries that have a similarity of 98.7% will have a distance of 1 (interval 99%-98% = distance 1). The default setting is 1%.

12. Leave the settings unaltered and press <***OK***>.

The *Advanced cluster analysis* window pops up. The *Network panel* displays the minimum spanning tree, the upper right panel (*Entry list*) displays the entries that are present in the tree. The *Selection entry list* lists the entries that are present in the selected node(s).

13. Select a node or branch by clicking on them. To select several nodes/branches hold the **Shift**-key while clicking.

As an exercise we will change some display settings.

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

15. In the *Node labels and sizes tab*, select ***Show node labels*** and select ***Dru Type*** from the ***Use label from*** list.

16. In the *Node colors tab*, select ***Number of entries*** from the drop-down list.

17. In the *Branch styles tab*, select ***branch length*** from the drop-down list.

18. In the *Branch labels and sizes tab*, select ***Show branch labels*** and ***branch length***.

19. Press <***OK***> to apply the new settings.

The *Advanced cluster analysis* window should now look like Figure 21.

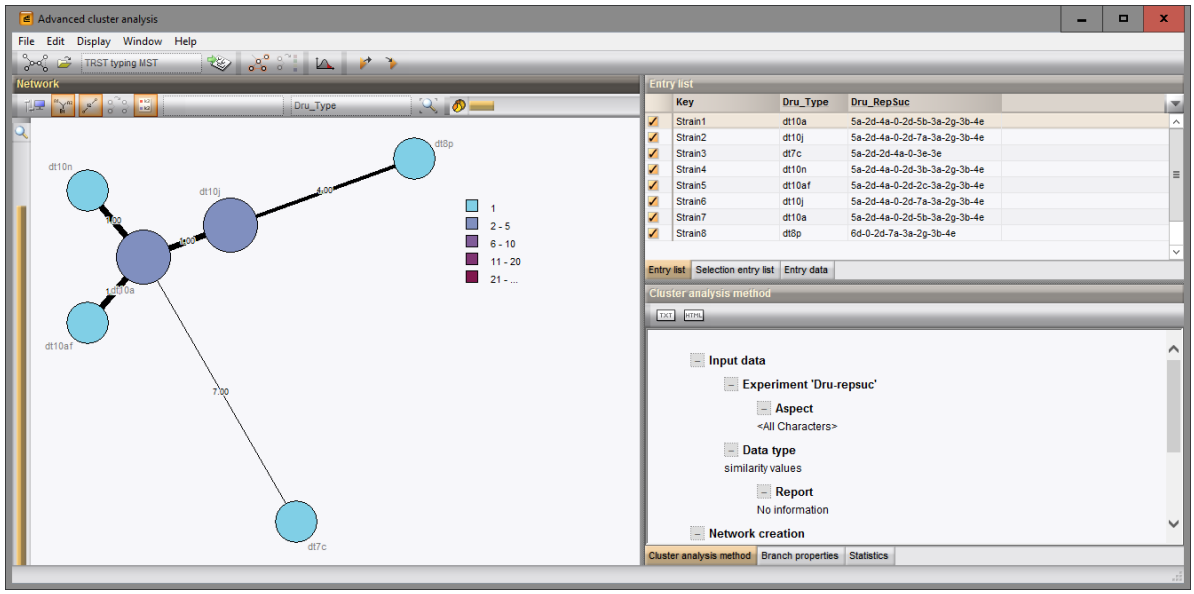

Figure 21: The *Advanced cluster analysis* window.

20. In the *Advanced cluster analysis* window, select ***Display*** > ***Zoom to fit*** or press to optimize the view of the tree in the current window.

21. Close the *Advanced cluster analysis* window.

8.4.3 UPGMA tree

Cluster analysis *sensu stricto* is based upon the similarity matrix and a subsequent algorithm for calculating bifurcating dendrograms to cluster the entries. In the *Polymorphic VNTR typing plugin*, you can choose between the following four methods: Unweighted Pair Group Method using Arithmetic averages (***UPGMA***), the ***Neighbor Joining*** method and two variants of UPGMA: ***Single linkage*** and ***Complete linkage***.

22. In the *Comparison* window, choose ***Repeat-Typing*** > ***Cluster types***.

23. Select ***UPGMA***, use the default alignment settings and default cost matrix and press <***OK***>.

The dendrogram is shown in the *Comparison* window.

24. Right-click in the header of the ***Dru-Type*** field in the *Information fields* panel and select ***Create groups from database field*** from the menu. Alternatively select ***Groups*** > ***Create groups from database field***.

25. Press <***Yes***> to create groups according to the assigned dru types.

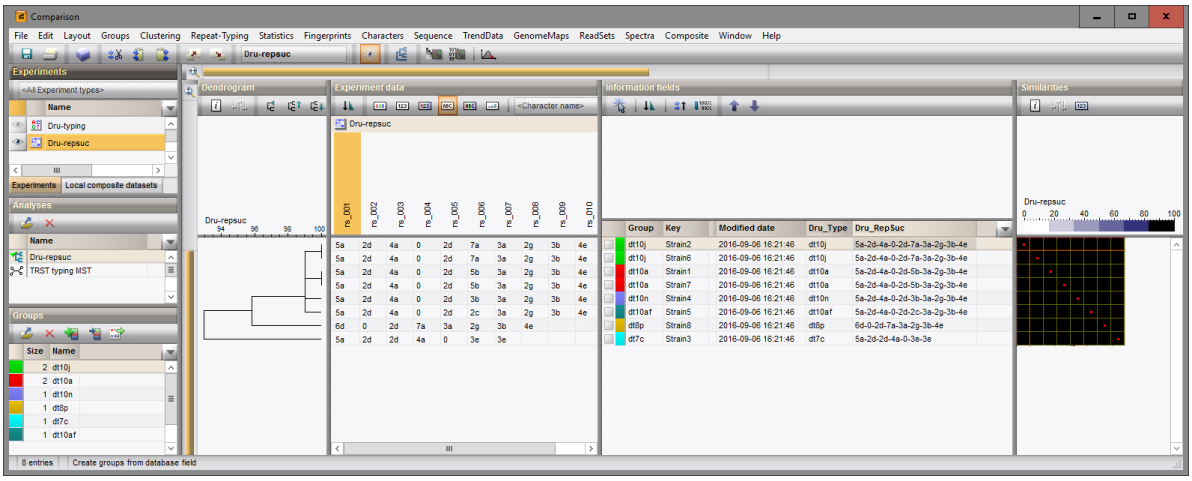

Figure 22: The *Comparison* window.

26. Click on the dendrogram to place a cursor on any node or tip (where a branch ends in an individual entry). The average similarity at the cursor's place is shown in the upper part of the *Experiment data* panel. You can move the cursor with the arrow keys.

27. Save and close the *Comparison* window.

Bibliography

[1] G. Benson. Sequence alignment with tandem duplication. *Journal of Computational Biology*, 4(3):351–367, 1997.