BioNumerics Tutorial:

Follow-up analysis of MLVA data

1 Aim

In this tutorial we will perform some cluster analyses on MLVA data. We will also see how we can alter the layout of the clusterings and how to export the pictures to use it in a publication, presentation, etc.

2 Preparing the database

1. Create a new database (see tutorial "Creating a new database") or open an existing database.

2. Import the MLVA repeat numbers from the example text file `MLVA repeat numbers.txt` as described in the tutorial: "Importing MLVA repeat numbers from a text file". This sample file contains repeat numbers for about 500 strains.

After import the *Main* window should look like Figure 1.

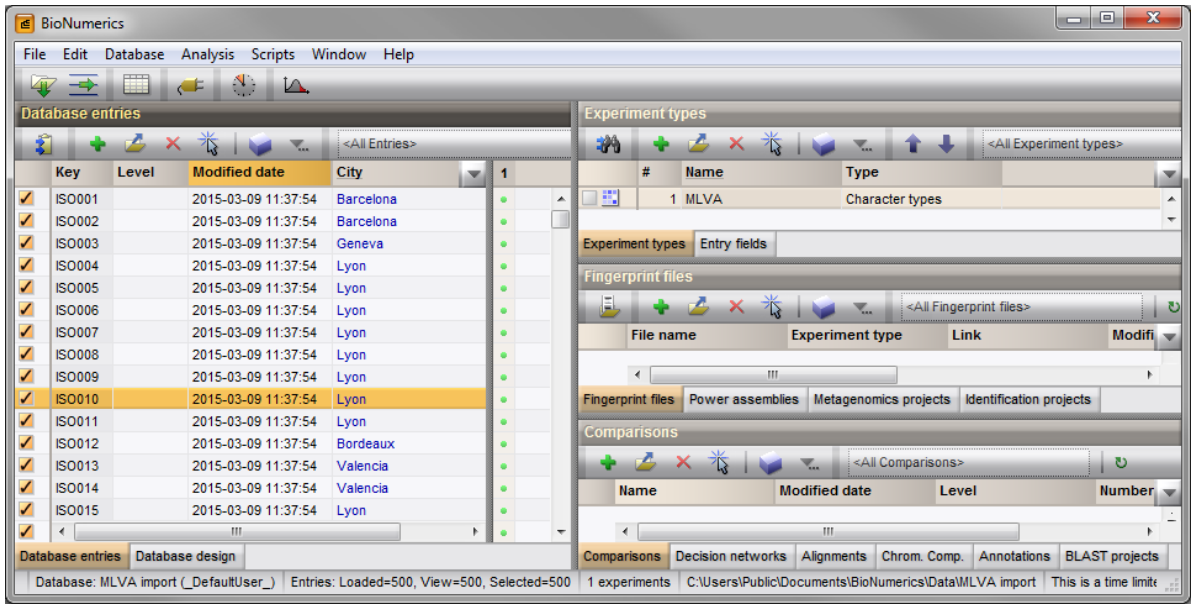

Figure 1: The *Main* window after import of the data.

The character data is stored in the character type **MLVA**.

3. To view the values in a list, double-click on the experiment **MLVA** in the *Experiment types* panel, select ***Settings*** > ***General settings...*** (), select the *Experiment card tab* and change the representation to ***List***. Close the two windows.

4. Click on a green colored dot in the *Experiment presence* panel to open the experiment card for an entry.

The imported repeat numbers are displayed in the experiment card next to the corresponding locus name (see Figure 2).

5. Close the experiment card by clicking in the left upper corner of the card.

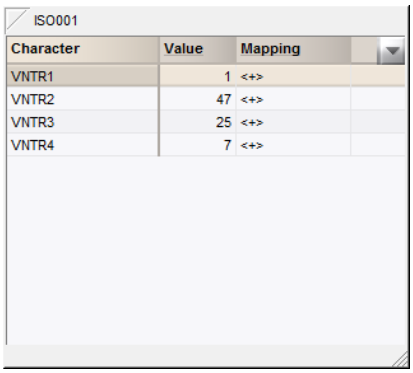

Figure 2: The experiment card.

3 Index of diversity

The Simpson's index of diversity measures the number and relative size of the different categories that are present in a character type experiment.

1. Select a few entries in the *Main* window.

2. Select ***Scripts*** > ***Browse internet...***, select ***Character related tools*** and ***Character index of diversity***.

A new dialog pops up, listing the number of selected entries and the character type experiments present in the database (see Figure 3).

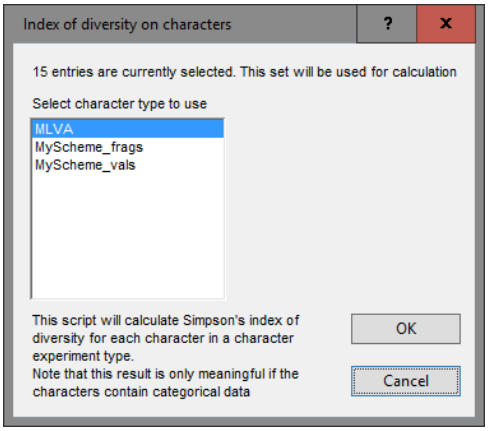

Figure 3: Select character type experiment.

3. Select the character type experiment containing the VNTR copy numbers and press <***OK***>.

The Simpson's index of diversity (S) is calculated for each character in the selected character type using the following formula:

$$S = 1 - \frac{\sum_{i=1}^{K} n_i(n_i - 1)}{N(N-1)}$$

where N is the number of selected entries, K the number of categories, and n_i the number of entries in category i.

The report displays for each VNTR in the selected character type, the ***Character*** name, the number of different categories detected in the selected entries for this character (***Count***), and the Simpson's index of

diversity (***Index***) (see Figure 4).

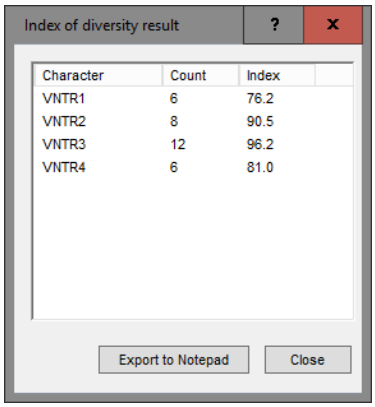

Figure 4: Report window.

The list can be sorted on the different columns by clicking the column headers. For the numerical columns, repeated clicking inverts the sorting.

The list can be exported as a tab-delineated file by pressing the <***Export to Notepad***> button. The export file, popped up as result.txt in Notepad, contains the character names, the counts, and the indices.

4. Close the report.

4 Creating a comparison

1. In the *Database entries* panel of the *Main* window, select all entries using ***Edit*** > ***Select all*** (**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. Click on the next to the experiment name **MLVA** in the *Experiments* panel and select ***Characters*** > ***Show values*** ( ) to display the repeat numbers in the *Experiment data* panel.

4. In the *Information fields* panel of the *Comparison* window, right-click in the header of the "City" field and select ***Create groups from database field*** from the floating menu. Alternatively select ***Groups*** > ***Create groups from database field***.

5. In the *Group creation preferences* dialog box press <***OK***> to create the comparison groups.

The groups appear in the *Groups* panel along with their color, size and name (see Figure 5).

The repeat numbers contained in the character set **MLVA** can be analyzed in BioNumerics with all the tools that are available to character data. This includes cluster analysis with a variety of methods and similarity coefficients. For VNTR repeat numbers, the coefficients that make most sense are:

- **Categorical coefficient**: preferred if differences in copy numbers should be treated in a qualitative way.

- **Euclidean distance**: preferred if differences in copy numbers should be treated in a quantitative way (larger difference means more distant organisms).

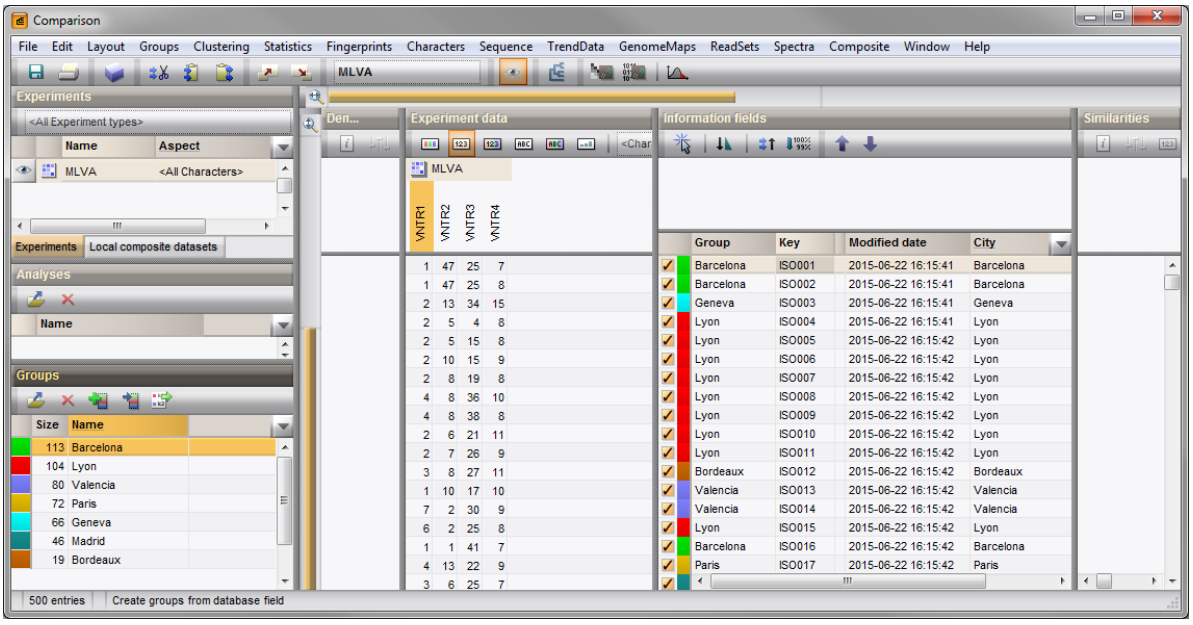

Figure 5: The *Comparison* window with comparison groups defined.

5 Creating a similarity based clustering

1. Make sure **MLVA** is selected in the *Experiments* panel and select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***.

The first step deals with the similarity coefficient for the calculation of the similarity matrix. The categorical coefficient compares the repeat numbers to see if they are the same or different but does not quantify the difference.

2. Select ***Categorical (values)*** from the list and press <***Next***>.

In step two the options related to the clustering algorithms are grouped. Under ***Method***, the clustering algorithm to be applied on the similarity matrix can be selected. A ***Dendrogram name*** can be entered in the corresponding text box. By default, the name of the experiment type appended with the aspect (here: "MLVA(<All characters>)") will be used.

3. Select ***UPGMA***, change the name of the analysis (e.g. **MLVA UPGMA Cat**) and <***Finish***> to start the cluster analysis.

During the calculations, the program shows the progress in the *Comparison* window's caption (as a percentage), and there is a green progress bar in the bottom of the window.

When finished, the dendrogram and the similarity matrix are displayed in their corresponding panels. The cluster analysis is listed in the *Analyses* panel of the *Comparison* window.

4. Press the **F4** key to clear any selection in the database.

5. Left-click on the dendrogram to place the cursor on any node or tip (where a branch ends in an individual entry).

6. To select entries in a cluster, click on the node of the cluster while holding the **Ctrl**-key.

7. Press ***Edit*** > ***Cut selection*** (, **Ctrl+X**) to remove the selected entries from the cluster analysis. Confirm the action. The dendrogram is automatically updated.

8. Select ***Edit*** > ***Paste selection*** (, **Ctrl+V**). The cluster analysis is recalculated automatically, and the selected entries are placed back in the dendrogram.

A branch can be moved up or down to improve the layout of a dendrogram:

9. Click the branch which you want to move up in the dendrogram and select ***Clustering*** > ***Move branch up*** ( ).

10. Click the branch which you want to move down in the dendrogram and select ***Clustering*** > ***Move branch down*** ( ).

To simplify the representation of large and complex dendrograms, it is possible to simplify branches by abridging them as a triangle.

11. Select a cluster of closely related entries and select ***Clustering*** > ***Collapse/expand branch*** ( ). Repeat this action to undo the abridge operation.

12. Select ***Clustering*** > ***Dendrogram display settings...*** ( ) to call the *Dendrogram display settings* dialog box.

13. Enable ***Show group colors*** and press <***OK***>.

The dendrogram branches are now colored according to the group colors (see Figure 6).

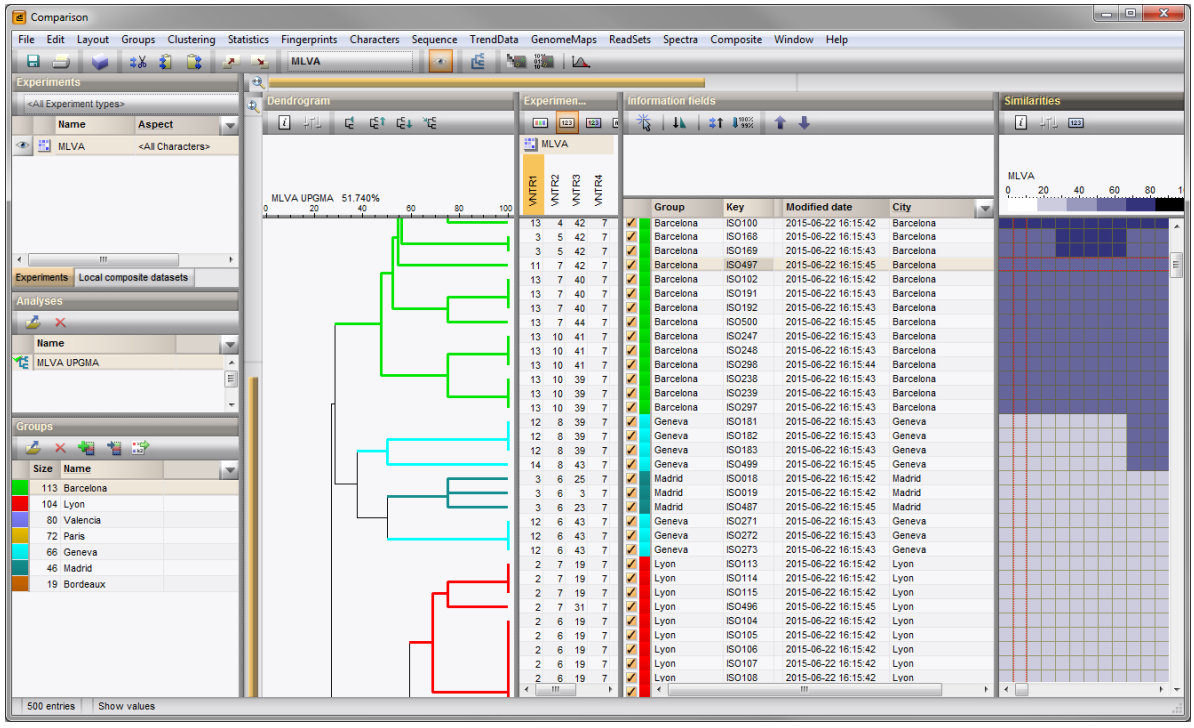

Figure 6: UPGMA dendrogram.

The similarity values in the *Similarities* panel are represented by shades of blue.

14. To show the values in the matrix, select ***Clustering*** > ***Similarity matrix*** > ***Show values*** ( ).

BioNumerics can export the cluster analysis as it appears in the *Comparison* window.

15. Select ***File*** > ***Print preview...*** (, **Ctrl+P**).

The *Comparison print preview* window now appears.

16. To scan through the pages that will be printed out, use ***Edit*** > ***Previous page*** (, **Page Up**) and ***Edit*** > ***Next page*** (, **Page Down**).

17. To zoom in or out, use ***Edit*** > ***Zoom in*** (, **Ctrl+Page Up**) and ***Edit*** > ***Zoom out*** (, **Ctrl+Page Down**) or use the zoom slider.

18. To enlarge or reduce the whole image, use ***Layout*** > ***Enlarge image size*** () or ***Layout*** > ***Reduce image size*** ().

19. If a similarity matrix is available, it can be included with ***Layout*** > ***Show similarity matrix*** ().

20. On top of the page, there are a number of small yellow slider bars, which can be moved.

21. To preview and print the image in full color select ***Layout*** > ***Use colors*** ().

22. Export the image to the clipboard with ***File*** > ***Copy page to clipboard*** () and selecting an appropriate format.

23. If a printer is available, use ***File*** > ***Print this page*** () or ***File*** > ***Print all pages*** () to print one or all pages.

24. Select ***File*** > ***Exit*** to close the *Comparison print preview* window.

25. Make sure **MLVA** is still selected in the *Experiments* panel and select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***.

26. Select ***Euclidean distance*** from the list and press <***Next***>.

27. Select ***UPGMA***, change the name of the analysis (e.g. **MLVA UPGMA Eucl**) and <***Finish***> to start the cluster analysis.

Both analyses are now listed in the *Analyses* panel. Switching between the different dendrograms can be done by simply double-clicking on the analysis name.

28. Save the comparison with the dendrograms by selecting ***File*** > ***Save*** ( , **Ctrl+S**). Specify a name (e.g. **All**) and press <***OK***>.

29. Close the saved comparison with ***File*** > ***Exit***.

6 Creating a minimum spanning tree

A minimum spanning tree in BioNumerics is calculated in the *Advanced cluster analysis* window. This window can be launched from the *Comparison* window.

1. Double-click on the saved comparison **All** in the *Comparisons* panel in the *Main* window.

2. Make sure ***MLVA*** is selected in the *Experiments* panel of the *Comparison* window.

3. Select ***Clustering*** > ***Calculate*** > ***Advanced cluster analysis...*** or press the  button and select ***Advanced cluster analysis*** to launch the *Create network* wizard.

The predefined template ***MST for categorical data*** uses the categorical coefficient for the calculation of the similarity matrix, and will calculate a standard minimum spanning tree with single and double locus variance priority rules.

4. Specify an analysis name (for example **MLVA1**), make sure ***MLVA*** is selected, select ***MST for categorical data***, and press <***Next***>.

To view and modify the settings of a selected template check the option ***Modify template settings for new analysis***.

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 *Cluster analysis method panel* displays the settings used, in this example the priority rules that result in the displayed network.

The colors of the comparison groups are automatically shown as node colors. The coloring can very easily be changed to a field state grouping defined in the *Main* window (not present here).

A node or branch can be selected by clicking on them. To select several nodes/branches hold the **Shift**-key, or click and hold down the left mouse button and drag the mouse pointer over the nodes to be selected.

5. Hold the **Shift**-key and select a few nodes.

The *Selection entry list panel* displays the entries currently selected on the network. The *Entry data panel* displays the character data for the selected entries.

6. The zoom slider on the left always further zooming in or out on the network. The zoom slider on top adjusts the size of the nodes.

7. Select ***Display*** > ***Zoom to fit*** or press  to optimize the view of the tree.

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

9. Uncheck the option ***Separate entries*** in the *Node colors tab*.

10. Check the option ***Show branch labels*** in the *Branch labels and sizes tab*.

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

The *Advanced cluster analysis* window should now look like Figure 7. The branch length - here the number of repeat number differences between the connecting nodes - is shown next to each branch.

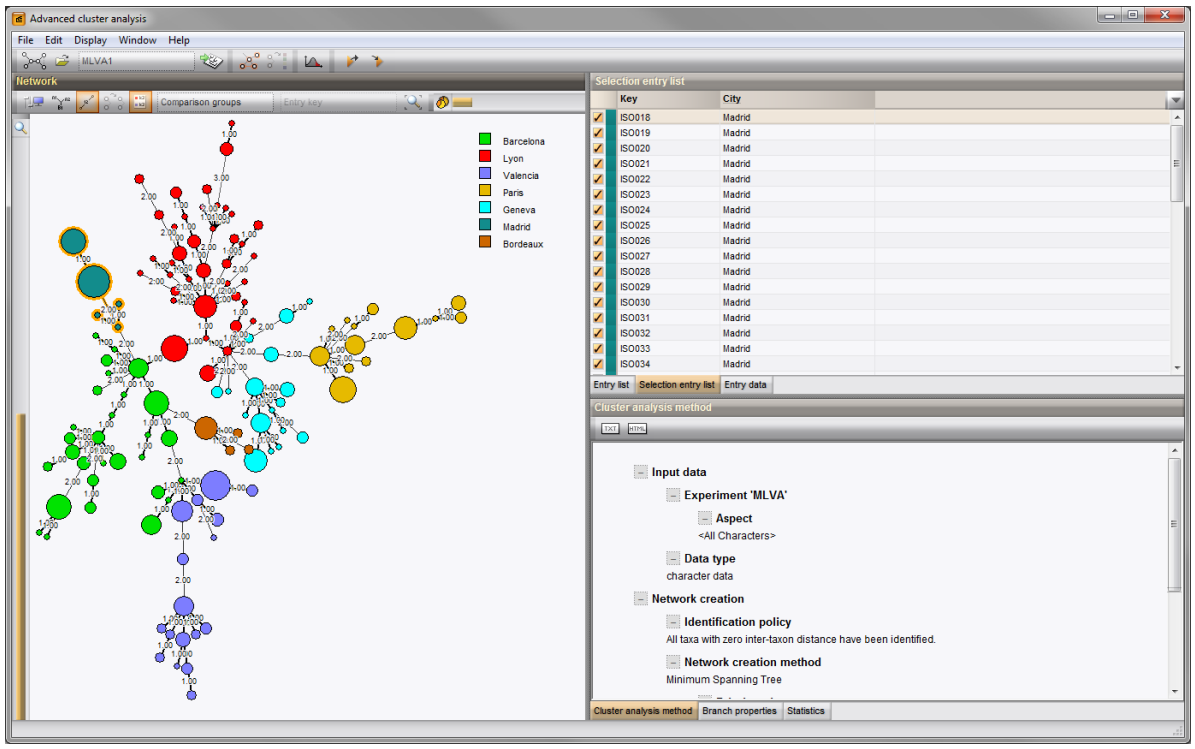

Figure 7: The *Advanced cluster analysis* window.

12. The image can be exported with ***File*** > ***Export image***.

13. Close the *Advanced cluster analysis* window and *Comparison* window with ***File*** > ***Exit***.