BioNumerics Tutorial:

Creating a custom mappings similarity matrix

1 Aim

In BioNumerics, character values can be mapped to categorical names according to predefined criteria (see tutorial "Importing non-numerical character data" for more information about the use of mappings in BioNumerics). When character mappings are present, it becomes possible to define a custom mappings similarity matrix, which determines how similarities are calculated among the mappings. This can be useful when analyzing data sets like SNPs, VNTRs, SSRs, etc. In this tutorial the use of a custom mappings matrix is illustrated.

2 Preparing the database

2.1 Introduction to the SNP demonstration database

The **SNP demonstration database** contains SNP data for 37 entries (see Figure 1) and can be downloaded directly from the *BioNumerics Startup* window (see 2.2), or restored from the back-up file available on our website (see 2.3).

10 SNPs were screened, and since the screening was performed on diploid organisms, 10 states (4 homozygous and 6 heterozygous states) are possible: A:A, C:C, G:G, T:T, C:A, G:A, T:A, G:C, T:C, and T:G.

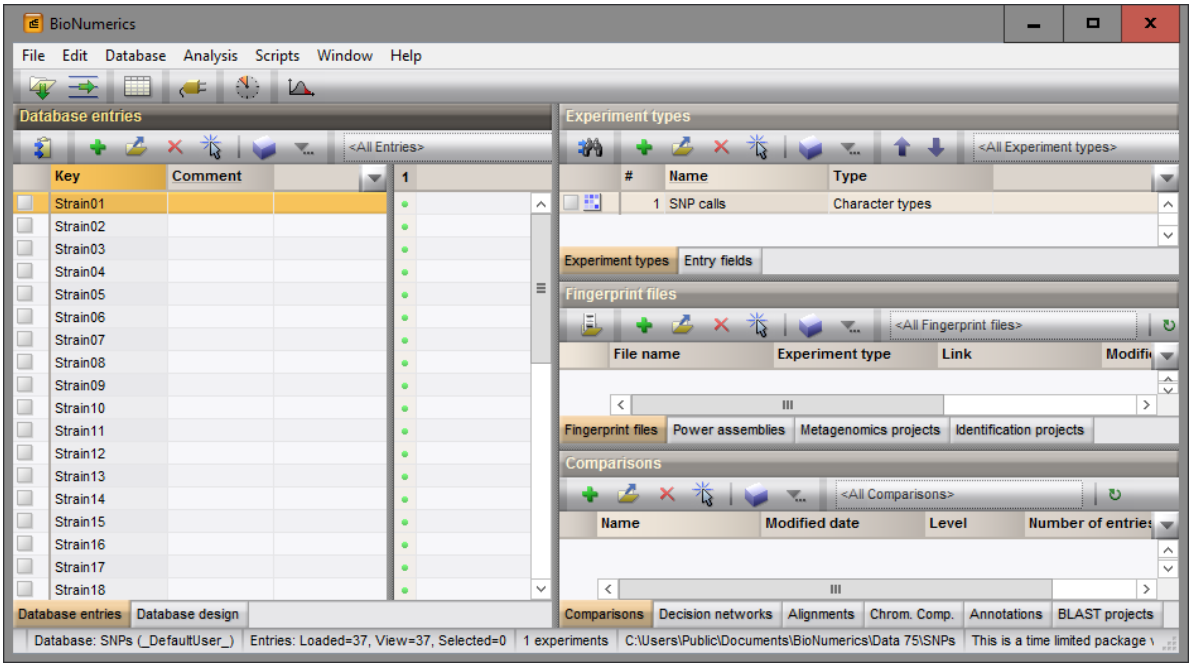


Figure 1: The *Main* window of the SNP demo database.

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 2).

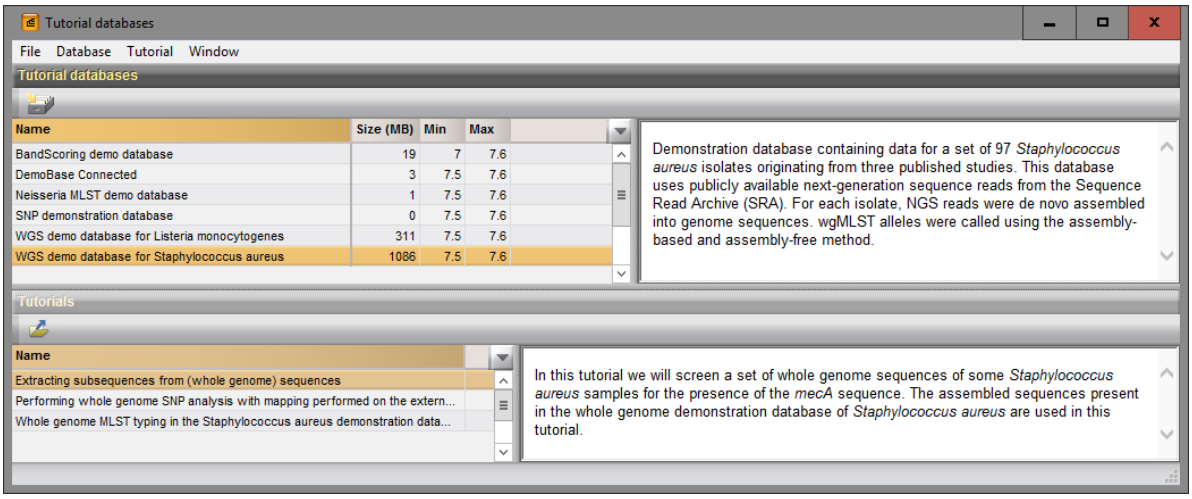


Figure 2: The *Tutorial databases* window, used to download the SNP demonstration database.

2. Select the **SNP demonstration database** 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 **SNP demonstration database** appears in the *BioNumerics Startup* window.

5. Double-click the **SNP demonstration database** in the *BioNumerics Startup* window to open the database.

The *Main* window should look like Figure 1.

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

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

6. Download the file `SNPs.bnbk` from `http://www.applied-maths.com/download/sample-data`, under 'SNP demonstration database'.

In contrast to other browsers, some versions of Internet Explorer rename the `SNPs.bnbk` database backup file into `SNPs.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. “SNP demonstration database”.

10. Click <***OK***> to start restoring the database from the backup file.

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

The *Main* window should look like Figure 1.

3 Character mappings

1. Double-click on the **SNP calls** character type experiment to open the *Character type* window.

The SNP set is displayed in the *Characters* panel (see Figure 3). For each SNP a character range is specified from 0 to 10, corresponding to the 10 different states (4 homozygous states, 6 heterozygous states).

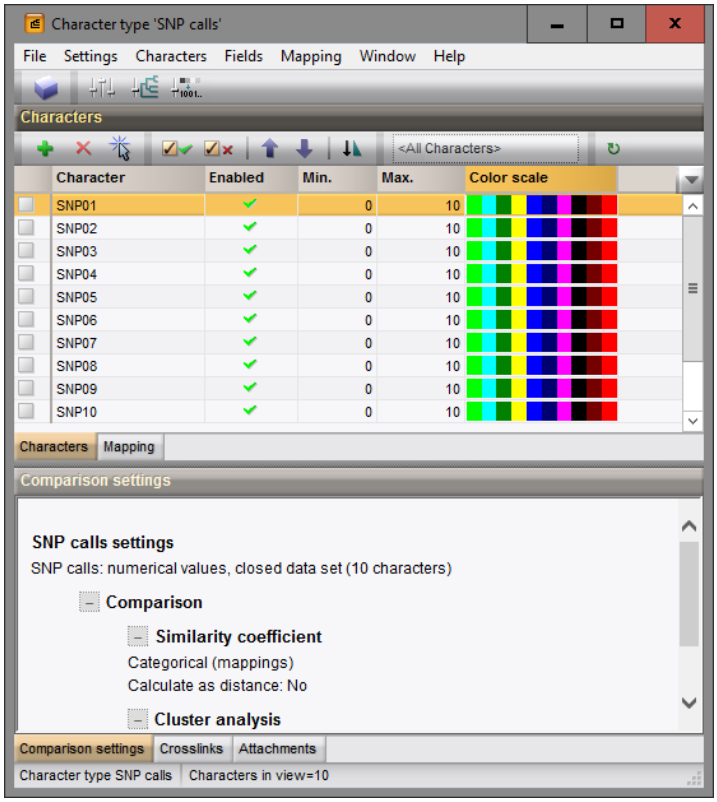


Figure 3: The SNP set.

2. Click on the *Mapping* panel to display the character mappings defined (see Figure 4).

The 10 possible states are mapped: A:A, C:C, G:G, T:T, C:A, G:A, T:A, G:C, T:C, and T:G.

3. Close the *Character type* window.

4. Click on a green colored dot in the *Experiment presence* panel to open the **SNP calls** experiment card for an entry (see Figure 5 for an example).

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

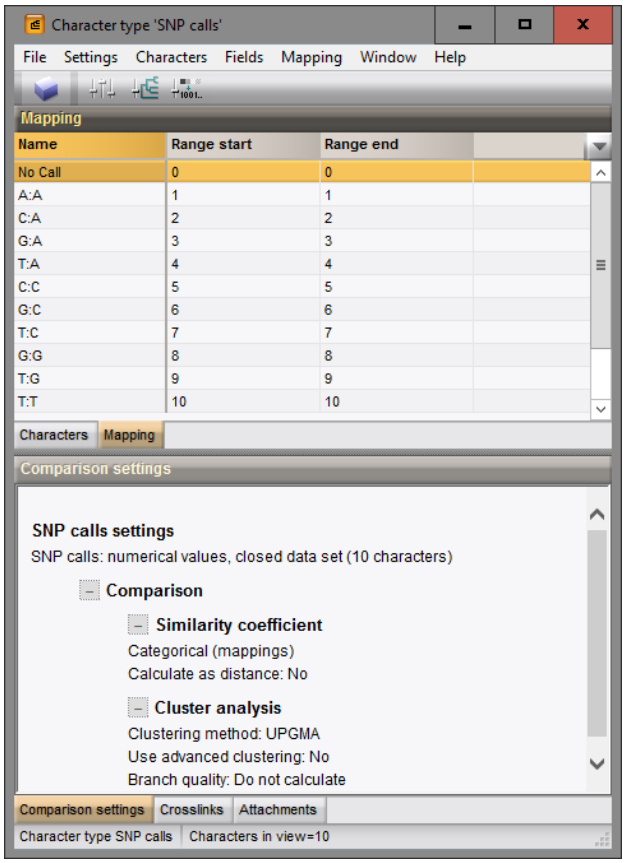


Figure 4: The character mappings.

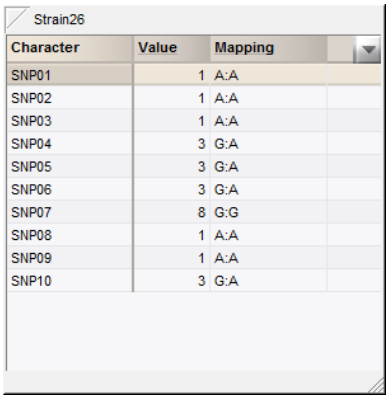


Figure 5: The character experiment card.

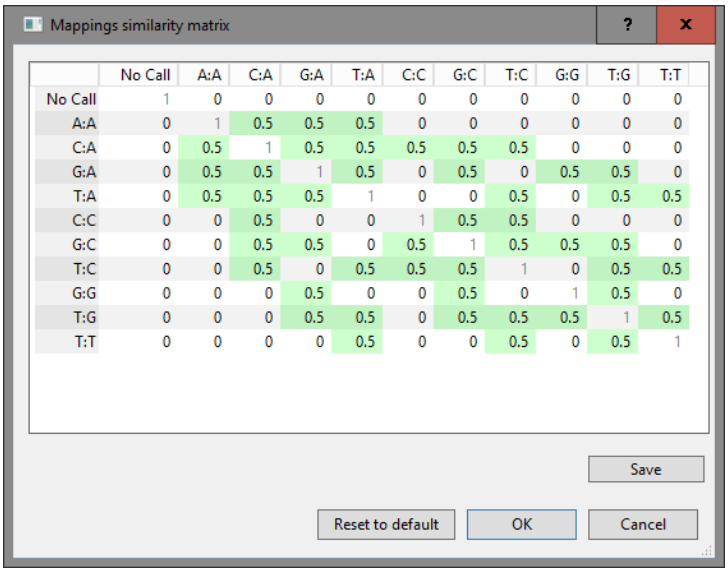
4 Custom mappings similarity matrix

When character mappings are present, it becomes possible to define a custom mappings similarity matrix, which determines how similarities are calculated among the mappings.

1. Double-click on the **SNP calls** character type experiment to open the *Character type* window.

2. Select ***Mapping*** > ***Edit mapping similarity matrix...*** to display the *Mappings similarity matrix* dialog box (see Figure 6).

The *Mappings similarity matrix* dialog box shows all defined character mappings in a matrix format. The default similarity value is ”1” or 100% match for self-matches (on the diagonal) and ”0” for matches between



	No Call	A:A	C:A	G:A	T:A	C:C	G:C	T:C	G:G	T:G	T:T
No Call	1	0	0	0	0	0	0	0	0	0	0
A:A	0	1	0.5	0.5	0.5	0	0	0	0	0	0
C:A	0	0.5	1	0.5	0.5	0.5	0.5	0.5	0	0	0
G:A	0	0.5	0.5	1	0.5	0	0.5	0	0.5	0.5	0
T:A	0	0.5	0.5	0.5	1	0	0	0.5	0	0.5	0.5
C:C	0	0	0.5	0	0	1	0.5	0.5	0	0	0
G:C	0	0	0.5	0.5	0	0.5	1	0.5	0.5	0.5	0
T:C	0	0	0.5	0	0.5	0.5	0.5	1	0	0.5	0.5
G:G	0	0	0	0.5	0	0	0.5	0	1	0.5	0
T:G	0	0	0	0.5	0.5	0	0.5	0.5	0.5	1	0.5
T:T	0	0	0	0	0.5	0	0	0.5	0	0.5	1

Figure 6: Custom similarity matrix.

different mappings. With other words, the default corresponds to a normal categorical matching. All values in the matrix, except for the ones on the diagonal, can be edited by clicking on the cell. Any value that deviates from the default ”0” will be highlighted in green. In the custom similarity matrix of this database, the states with a difference in only one allele, have a 50% match (”0.5”) (see Figure 6). The mappings similarity matrix will be used when similarities are calculated with the ***Categorical (mappings)*** coefficient in the *Comparison* window (see 6).

3. Close the *Mappings similarity matrix* dialog box and the *Character type* window.

5 Comparison window

1. In the *Database entries* panel of the *Main* window, select all entries with the keyboard shortcut **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.

All 37 entries are loaded in the *Comparison* window.

3. Click on the  next to the experiment name **SNP calls** in the *Experiments* panel to display the data in the *Experiment data* panel.

Initially, the character values are displayed as colors according to the color scale defined for each character (see Figure 3).

4. Select ***Characters*** > ***Show mappings+colors*** () to show the mappings in overlay with the colors (see Figure 7).

6 Cluster analysis

Cluster analysis is a two-step process. First, all pairwise similarity values are calculated with a **similarity coefficient**. Then, the resulting similarity matrix is converted into a dendrogram with a **clustering algorithm**. Although in practice these steps are performed together, they each require their own comparison

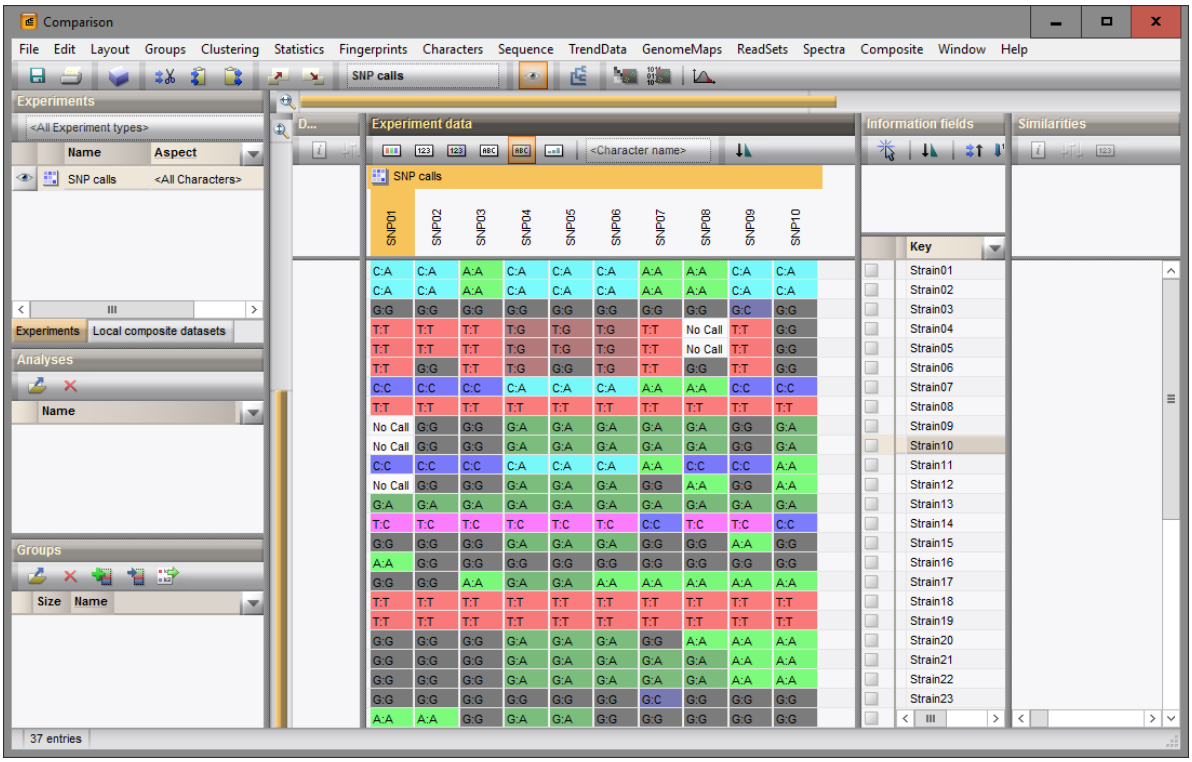


Figure 7: The *Comparison* window.

settings.

1. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...*** in the *Comparison* window.

The first step deals with the similarity coefficient for the calculation of the similarity matrix.

2. Select ***Categorical (mappings)*** from the list and press <***Next***> (see Figure 8).

The mappings similarity matrix will be used when similarities are calculated with the ***Categorical (mappings)*** coefficient.

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: ”SNP calls (<All characters>)”) will be used.

3. Select ***UPGMA*** and <***Finish***> to start the cluster analysis.

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 (see Figure 9).

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

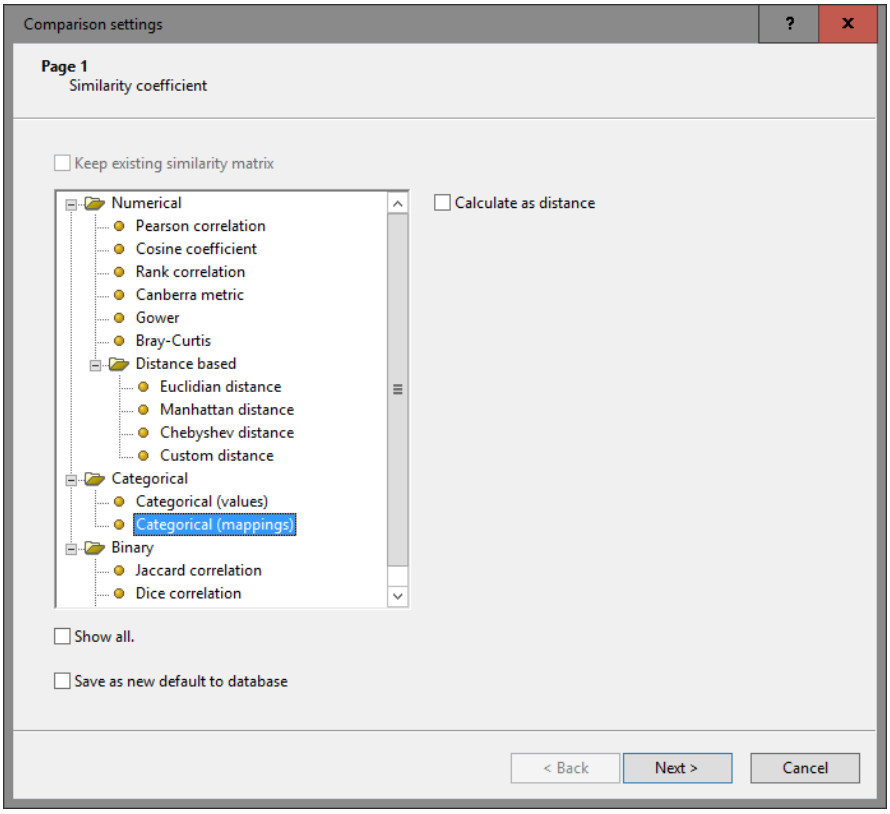


Figure 8: Cluster analysis: step 1.

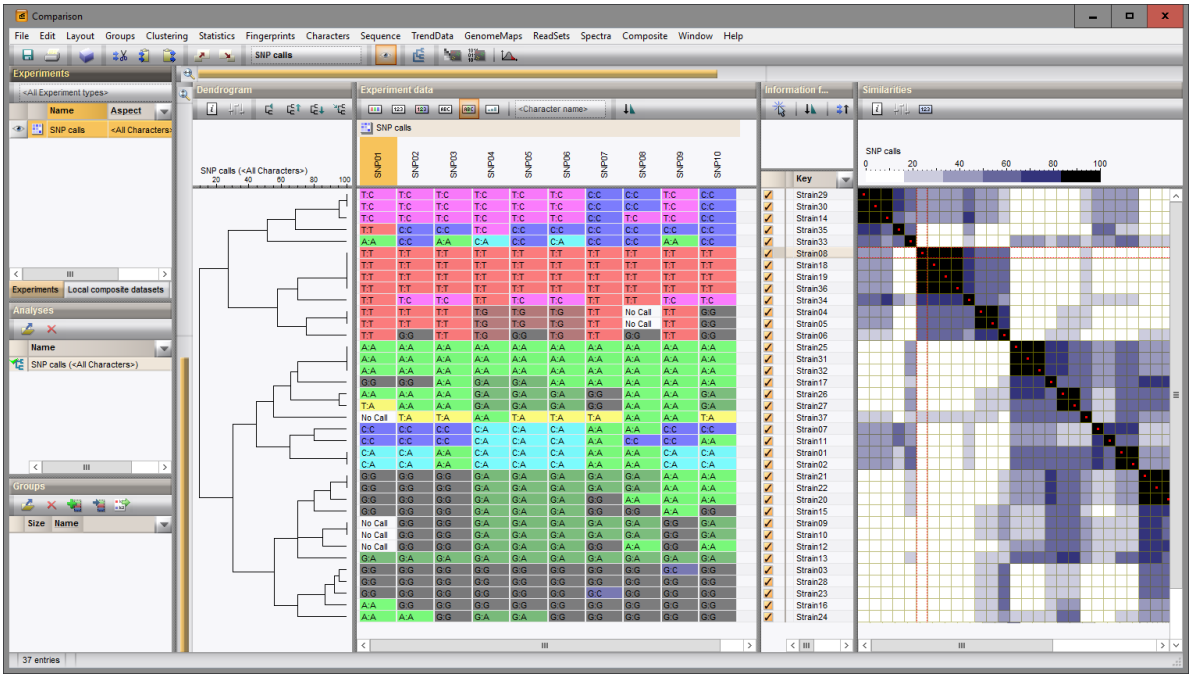


Figure 9: The *Comparison* window.