

# GeneMaths XT

## From Spots to Genes

---

---

**Copyright © 1998, 2007, Applied Maths NV. All rights reserved.**

GeneMaths XT is a registered trademarks of Applied Maths NV.  
All other product names or trademarks are the property of their respective owners.



[www.applied-maths.com](http://www.applied-maths.com)



# NOTES

---

## SSUPPORT BY APPLIED MATHS

While the best efforts have been made in preparing this manuscript, no liability is assumed by the authors with respect to the use of the information provided.

Applied Maths will provide support to research laboratories in developing new and highly specialized applications, as well as to diagnostic laboratories where speed, efficiency and continuity are of primary importance. Our software thanks its current status for a part to the response of many customers worldwide. Please contact us if you have any problems or questions concerning the use of GeneMaths XT, or suggestions for improvement, refinement or extension of the software to your specific applications:

**Applied Maths NV**  
**Keistraat 120**  
**9830 Sint-Martens-Latem**  
**Belgium**  
PHONE: +32 9 2222 100  
FAX: +32 9 2222 102  
E-MAIL: [info@applied-maths.com](mailto:info@applied-maths.com)

**Applied Maths, Inc.**  
**13809 Research Boulevard, Suite 645**  
**Austin, Texas 78750**  
**U.S.A.**  
PHONE: +1 512-482-9700  
FAX: +1 512-482-9708  
E-MAIL: [info-US@applied-maths.com](mailto:info-US@applied-maths.com)

URL: [www.applied-maths.com](http://www.applied-maths.com)

## LIMITATIONS ON USE

The GeneMaths XT software and this accompanying guide are subject to the terms and conditions outlined in the License Agreement. The support, entitlement to upgrades and the right to use the software automatically terminate if the user fails to comply with any of the statements of the License Agreement.

No part of this guide may be reproduced by any means without prior written permission of the authors.

**Copyright © 1998, 2007, Applied Maths NV. All rights reserved.**

GeneMaths XT is a registered trademark of Applied Maths NV.  
All other product names or trademarks are the property of their respective owners.



# Table of contents

---

<b>1. Import</b> .....	<b>5</b>	4.4 Arrange columns.....	19
1.1 Introduction.....	5	4.5 Profile .....	20
1.2 Importing the data.....	5		
1.3 Some features of the Main View .....	7		
<b>2. Annotation</b> .....	<b>9</b>		
<b>3. Column groupings</b> .....	<b>11</b>		
<b>4. Preprocessing</b> .....	<b>15</b>	<b>5. Statistics &amp; Analysis</b> .....	<b>21</b>
4.1 Subset .....	15	5.1 Statistic tests on the column groups.....	21
4.2 Preprocessing diagram.....	15	5.2 Remove effect.....	23
4.3 From spots to genes.....	19	5.3 Pattern matching.....	23
		5.4 Venn diagram .....	24
		5.5 Row grouping .....	25
		5.6 Hierarchical clustering.....	26
		5.7 Dimensioning techniques.....	27
		5.8 Partitioning.....	28
		5.9 Self-Organizing map.....	29



# 1. Import

## 1.1 Introduction

The data files used in this quickguide are GenePix files. These files are installed with the software in the `\My GeneMaths Session` directory. The dataset is a fictitious dataset and is only used for illustration purposes. It is not the aim to discuss any scientific context.

## 1.2 Importing the data

1.2.1 Start GeneMaths XT by double clicking on the



icon on the desktop or select *Start > Programs > Applied Maths > GeneMaths XT* from the taskbar.

1.2.2 Click *<Next>* in the welcome screen to begin the import of the data. The *Import* wizard pops.

1.2.3 Select the second option *Import a set of files each containing one array* and hit *<Next>* (see Figure 1-1).

In the next step, predefined formats are listed. A **format** defines the source and the content of the data. If your source is not listed here, you can define your own custom format. In this exercise, we are going to use the predefined **Genepix** format.

1.2.4 Select **Genepix** from the format list. A short description of the format is shown in the right panel (Figure 1-2). Hit *<Next>*.

In the next step of the import wizard, some predefined mappings are listed. A **mapping** specifies which quantitations and quality control items will be imported in the

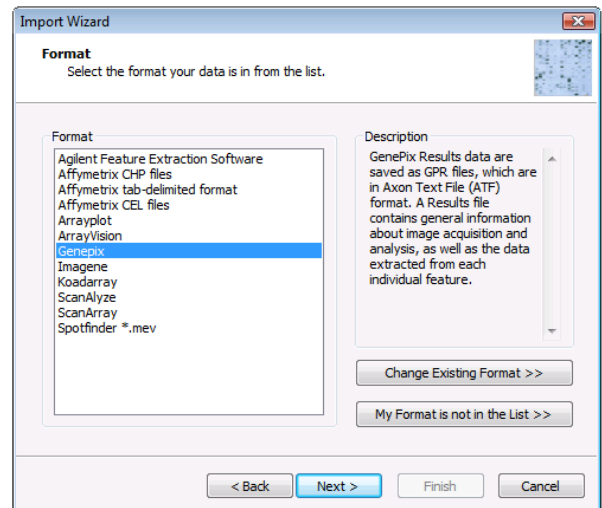


Figure 1-2. Select format.

session. In this exercise, we are going to use the mapping **Raw median**.

1.2.5 Make sure the **Raw median** mapping is selected and press *<Next>* (see Figure 1-3).

1.2.6 In the next window, select *Import array files (without extra biological information)* and press *<Next>* (see Figure 1-4).

1.2.7 In the next window, navigate to the path where the .gpr files are stored. Select all files and press *<Open>* (see Figure 1-5).

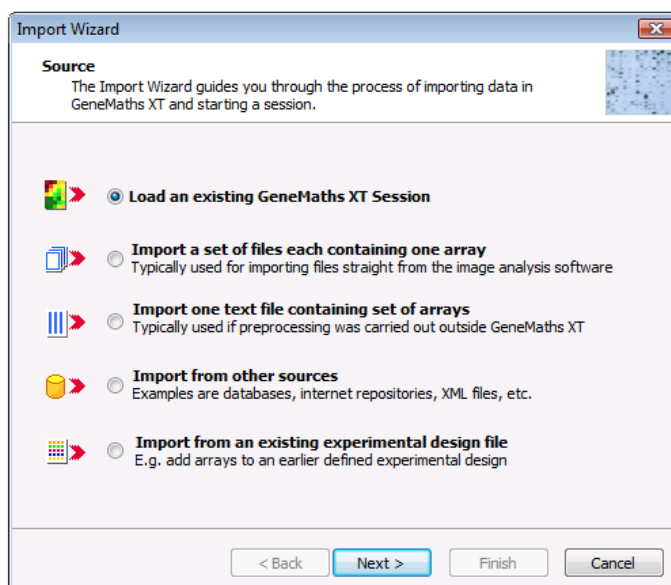


Figure 1-1. Import wizard: Select data source.

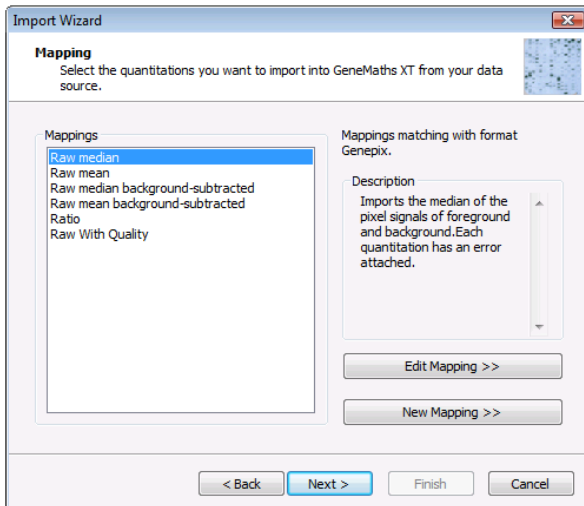


Figure 1-3. Select mapping.

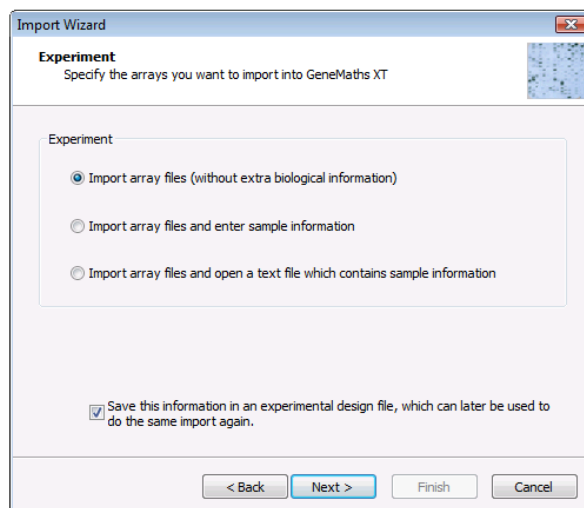


Figure 1-4. Specify the arrays you want to import

1.2.8 GeneMaths XT now asks you to save the experimental design file as a .XPS file. Specify a name e.g. **Guided Tour.xps** and press the **<Save>** button.

The **experimental design** file contains the link between the data and the array information. Depending on your design this can be really simple or can become more complex when using replicates and colorflips.

After the processing of the data, GeneMaths XT will open a session with 7 layers (see Figure 1-6). Select the first layer **Target**. The expression matrix is updated.

The session (see Figure 1-7) contains six row identifiers and one column identifier (see Figure 1-8 and Figure 1-9).

1.2.9 Save this session with **File > Save Session**.

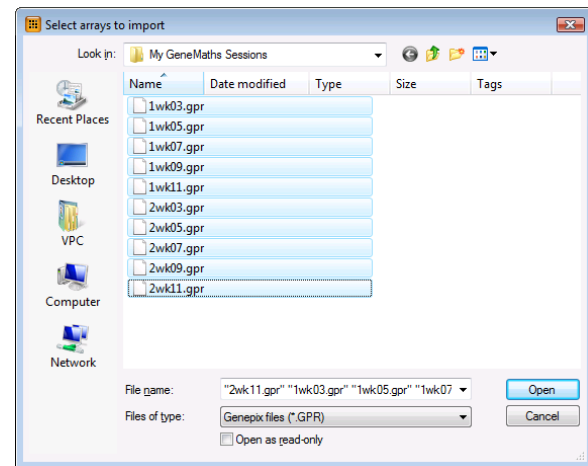


Figure 1-5. Start Wizard window: Data files

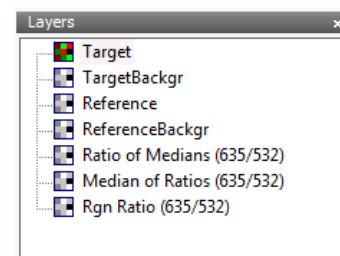


Figure 1-6. Seven layers are imported.

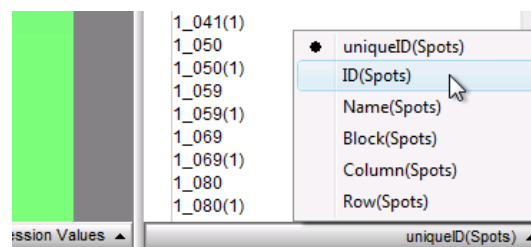


Figure 1-8. Six row identifiers.

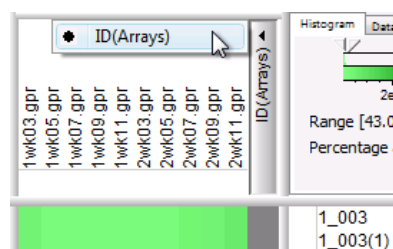


Figure 1-9. One column identifier.

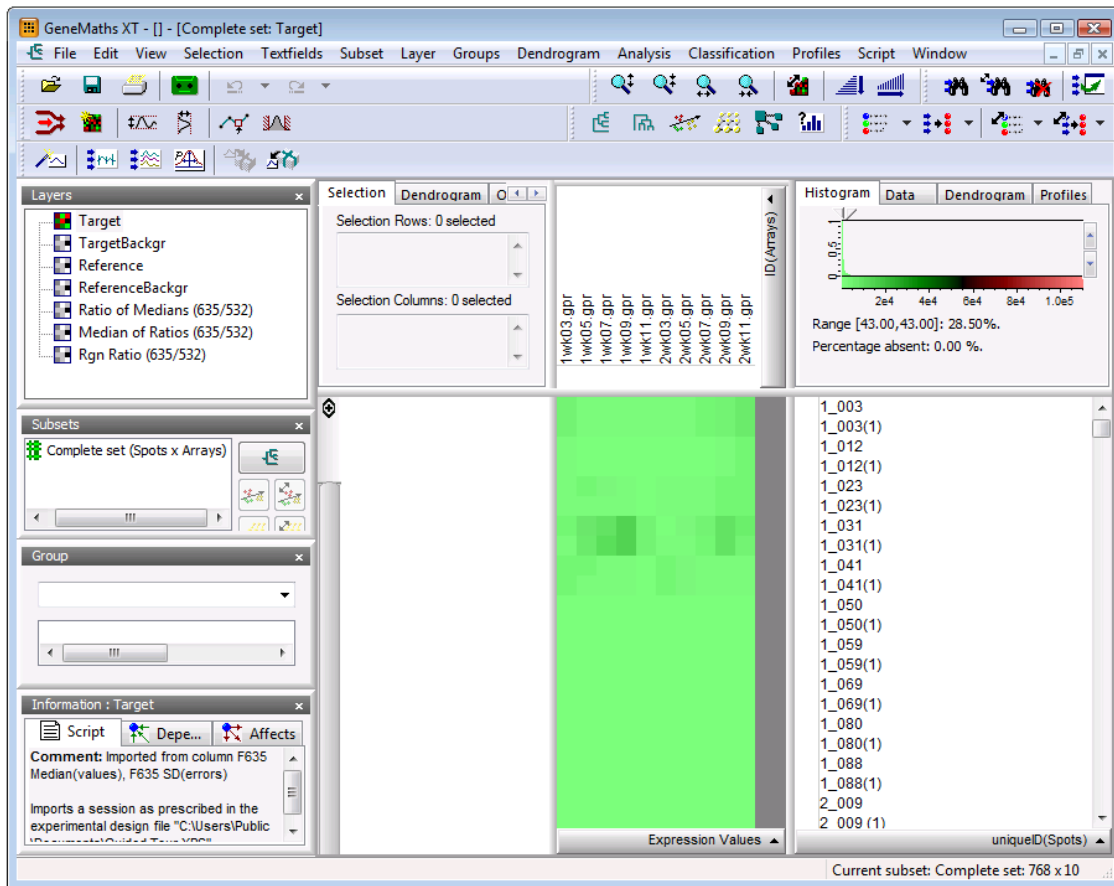


Figure 1-7. The Main window after import.

### 1.3 Some features of the Main View

The data set panel shows the expression matrix in color code. The color palette can be changed if desired.

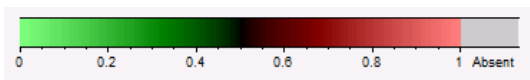


Figure 1-10. Default color palette.

If you hover over the expression matrix with your mouse, an information box for the underlying data cell is shown with the value, error and selected layer.

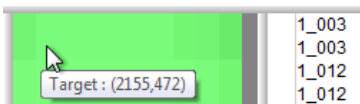


Figure 1-11. Information box.

1.3.1 Use the horizontal zoomslider to zoom in on the matrix. The expression values are shown (see Figure 1-12.).

Panels with information fields for the rows and columns in the matrix are located next to the data set panel ,

horizontal zoom slider

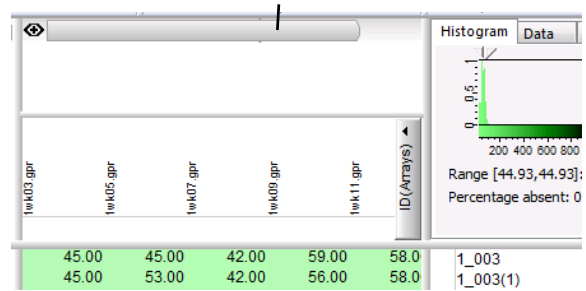


Figure 1-12. Horizontal zoom slider.

together with panels where the dendrograms and profiles are visualised when constructed.

The info panels in the upper corners of the data set panel contain information on the dendrograms, current selection, profiles, histogram of the values in the expression matrix, ....

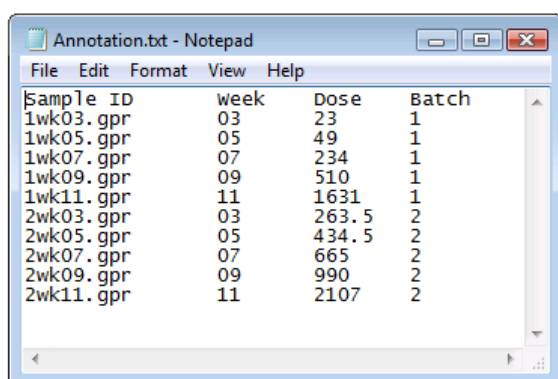
The Main View contains a number of dockable windows (see Figure 1-7):

- **Layers window:** contains the list of layers present in the session. The layers are structured in a tree indicating the dependencies.

- **Subsets window:** contains the list of subsets and scopes in the session, structured in a tree indicating the dependencies.
- **Information window:** shows the active history of layers, subsets and scopes in the session.
- **Group window:** lists the groupings and groups defined in the session.

## 2. Annotation

Additional array information is present in a tab-delimited text file called **Annotation.txt**. This text file can be found in the **\My GeneMaths Sessions** directory.



Sample ID	week	Dose	Batch
1wk03.gpr	03	23	1
1wk05.gpr	05	49	1
1wk07.gpr	07	234	1
1wk09.gpr	09	510	1
1wk11.gpr	11	1631	1
2wk03.gpr	03	263.5	2
2wk05.gpr	05	434.5	2
2wk07.gpr	07	665	2
2wk09.gpr	09	990	2
2wk11.gpr	11	2107	2

Figure 2-1. Text file containing array information.

We are going to import this additional information in our GeneMaths XT session.

2.0.1 Select **Textfields > Import** in the *Main* window of GeneMaths XT. The *Import Text Fields* dialog box pops up (see Figure 2-2).

2.0.2 Select the aspect **Arrays** and navigate to the data folder where the text file is stored. Select the text file.

In order to link the entries from the text file to the entries present in the session we need to select the field in the session and the column in the text file that contain exactly the same entries:

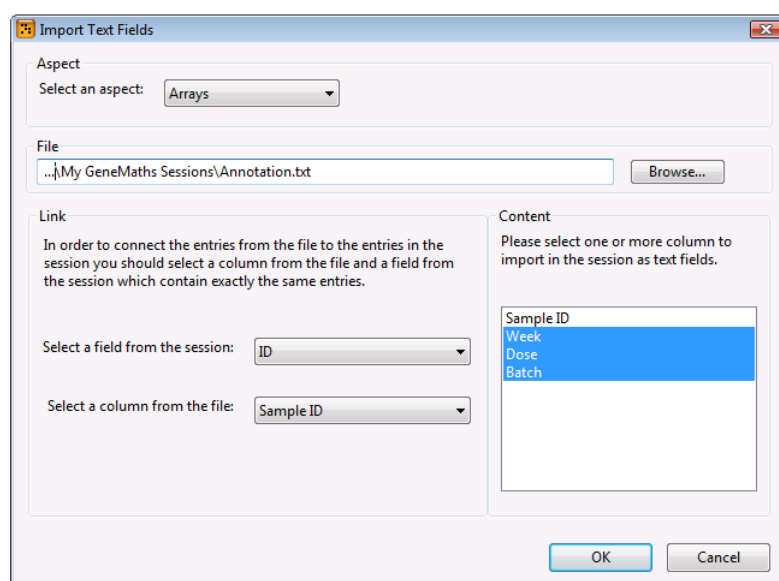


Figure 2-2. The *Import Text Fields* dialog box.

- *Select a field from the session:* select **ID**.

- *Select a column from the file:* select **Sample ID**.

Next, we need to select the column(s) from our text file that we want to import in our session.

2.0.3 Select the **Week**, **Dose**, and **Batch** columns in the *Content* panel (see Figure 2-2). Use the CTRL-button to select more than one column from the list. Press **<OK>**.

2.0.4 Click on the column identifier tab in the main window of GeneMaths XT. The information fields are added to the list of identifiers (see Figure 2-3).

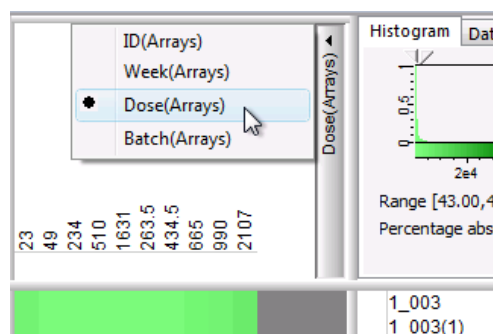


Figure 2-3. Three new column identifiers.

2.0.5 Select one of the new column identifiers from the list, e.g. **Dose**.

2.0.6 The values are updated in the *Column names* panel (see Figure 2-3).



## 3. Column groupings

In some cases it is known in advance that the data set contains separate groups while in other cases one will try to find out by means of an analysis what the groups in the data set are.

Groups are used as input for some analyses and statistic tests and are also used as a visual aid tool. Managing group appearance (i.e. colors, names and symbols) for row or column entries can be done from all views.

In our example session, we are going to make groups based on the **Batch** and **Week** values.

### • 'Batch' grouping

In the first step, we are going to make column groups based on the **Batch** column identifier (see Figure 3-1).

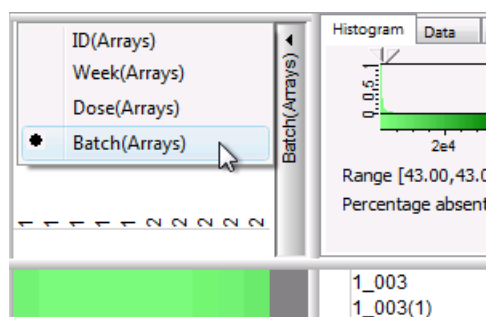


Figure 3-1. The 'Batch' identifier.

3.0.1 Select **Groups > Edit Column Groups** and click on **<Create New Grouping>**.

3.0.2 In the next window, select **Batch** from the drop-down menu and press **<OK>**.

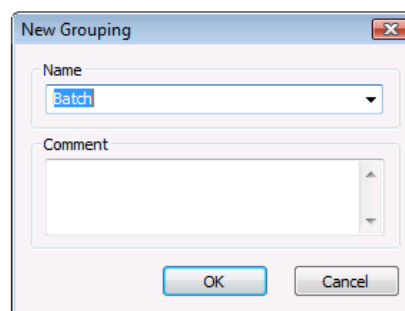


Figure 3-2. Batch grouping

3.0.3 In the next window, uncheck all checkboxes and press **<OK>** (see Figure 3-3).

The groups based on the settings are shown in the next window (see Figure 3-4).

3.0.4 Press **<Exit>**.

3.0.5 Select **Batch** from the list of column identifiers (if not already done).

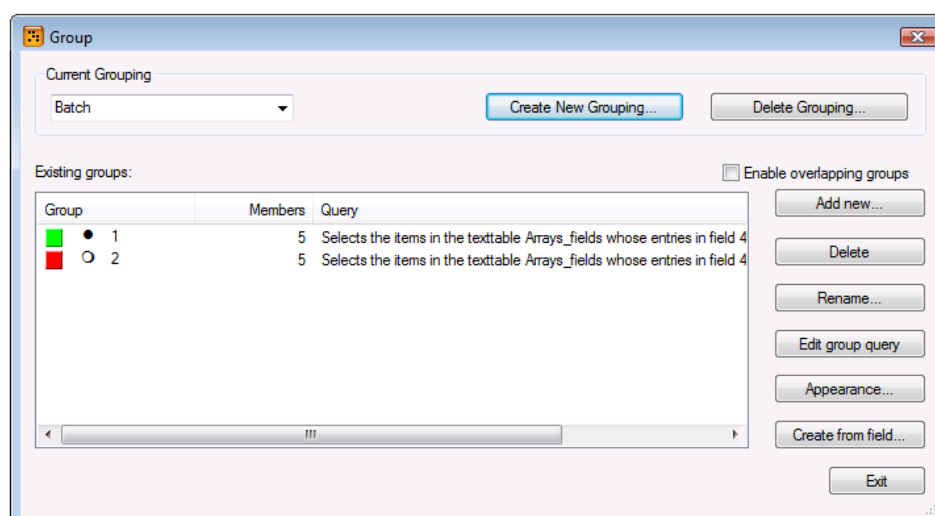


Figure 3-4. The *Group* dialog box: groups based on the batch number.

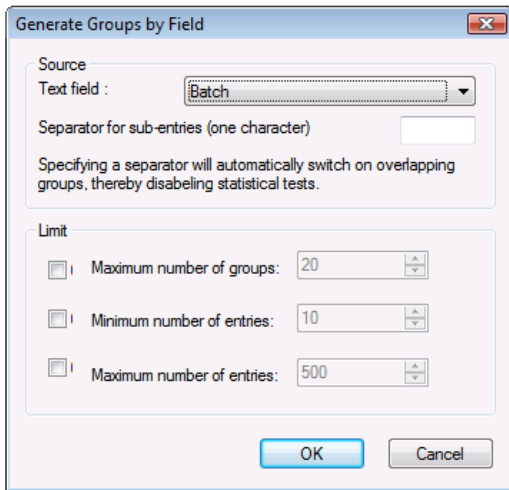


Figure 3-3. Group settings.

The groups are visualized by different colors in the *Column names* panel:

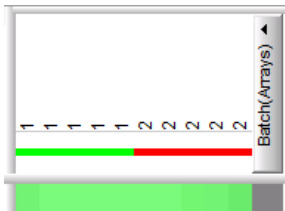


Figure 3-5. Groups based on the batch number.

•‘Week no’ grouping

Now we are going to make column groups based on the **Week** column identifier.

3.0.6 Select *Groups > Edit Column Groups* and click on *<Create New Grouping>*.

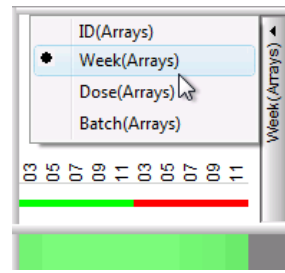


Figure 3-6. The ‘Week’ column identifier.

3.0.7 In the next window, select **Week** from the drop-down menu and press *<OK>*.

3.0.8 In the next window, uncheck all checkboxes and press *<OK>*.

The groups based on the settings are shown in the next window (see Figure 3-7).

3.0.9 Press *<Exit>*.

3.0.10 Select **Week** from the list of column identifiers (if not already done).

The groups are visualized by different colors in the *Column names* panel:

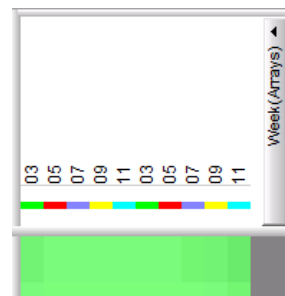


Figure 3-8. Groups based on the Week number.

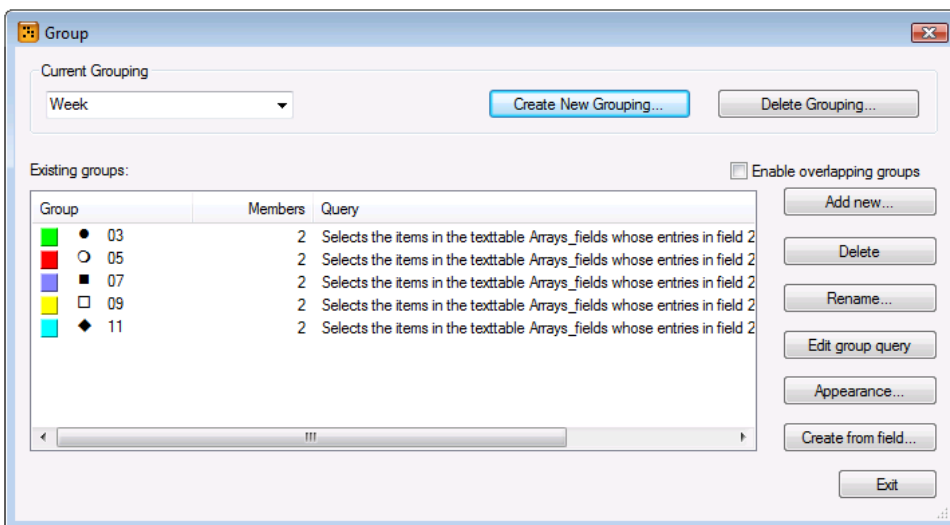



Figure 3-7. The *Group* dialog box: groups based on the week number.

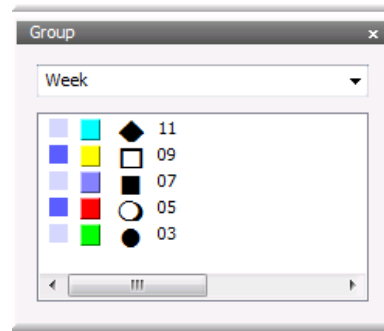
To display the colors from another grouping press the arrow next to the  button and select the grouping from the list.

In the *Group* window, the created groupings are listed in the drop-down menu.

3.0.11 Select **Week** in the *Group* window (see Figure 3-9).

The groups based on the week number are listed below the drop-down menu (see Figure 3-9). Group members can be selected from within this panel.

3.0.12 Select all members belonging to a group by CTRL-clicking on the square next to the group color. The square is highlighted in blue and the member of the group are selected in the *Array* panel.



**Figure 3-9. The *Group* window .**

3.0.13 Select another group while holding the CTRL-button (see Figure 3-9).

3.0.14 Use the SHIFT-button to select a range of groups.

To unselect all arrays press the F4 button.



## 4. Preprocessing

### 4.1 Subset

4.1.1 Select the **ID** row identifier from the list of row identifiers. The row names in the *Row names* panel are updated (see Figure 4-1).

4.1.2 Scroll through the row entries. Some rows do not correspond to spots of real genes:

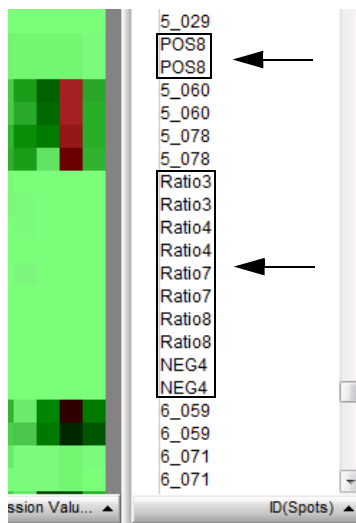


Figure 4-1. Some rows do not correspond to spots.

In a next step we are going to create a **subset**, containing all rows that correspond to spots of real genes. We are going to leave those rows out that do not correspond to spots of genes. We are going to make a selection in our session, with the use of a query:

4.1.3 Launch the *Row Query* tool with *Selection > Row Selection from Query* or press CTRL+Q.

4.1.4 Enter in the *Text* box in the lower panel of the window **NEG, POS, CAL, Ratio**, separated by a pipe (|) delimiter (see Figure 4-2).

4.1.5 Check *Use Regular Expressions*. Make sure the **ID** field is selected and the logic *Replace current selection* option is checked. Press <OK>.

The row entries, having the text **NEG, POS, CAL** or **Ratio** in their ID information field are selected.

4.1.6 Select *Selection > Invert Row Selection*.

The selection is inverted: 720 entries that correspond to spots of genes are selected.

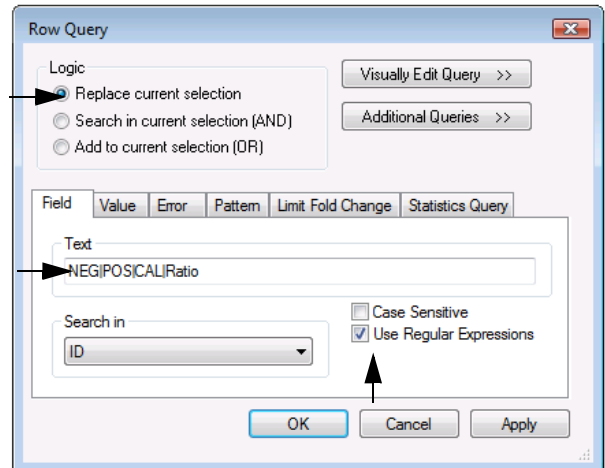


Figure 4-2. Row Query dialog box.

4.1.7 Select *Subset > Selection to Subset*. Give a name to the subset e.g. **OnlyGenes** and press <OK>.

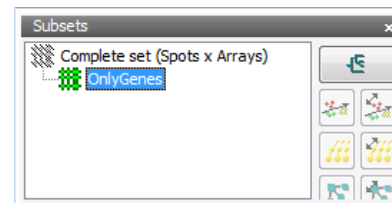


Figure 4-3. The Subsets window.

The new subset is listed in the *Subsets* window as a child of the Complete set.

4.1.8 Press F4 to clear the current selection.

### 4.2 Preprocessing diagram

In order to perform data analysis on this dataset, we first need to preprocess our data.

4.2.1 Make sure the **OnlyGenes** subset is selected in the *Subsets* window. This subset will be set as default subset parameter when selecting a preprocessing tool.

4.2.2 In GeneMaths XT, select *Layer > Preprocessing diagram*.

In the *Preprocessing* window, all layers present in the session are displayed. On the right side of the window, the preprocessing tools are listed.

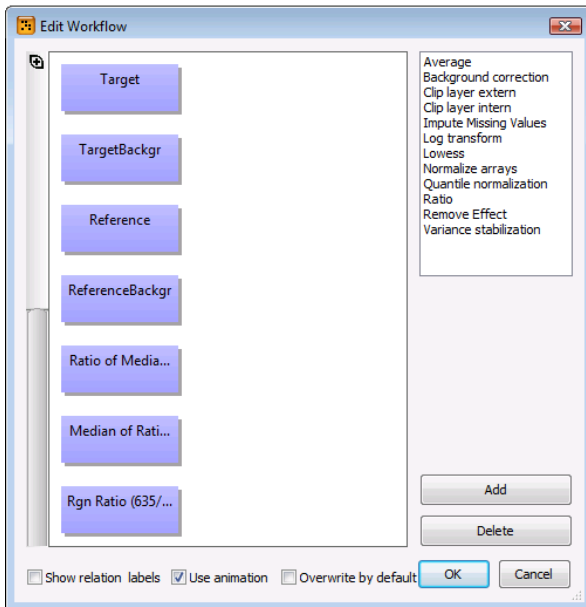


Figure 4-4. *Preprocessing window.*

### • Background correction

4.2.3 In the *Preprocessing* window, select the **Target** and **TargetBackgr** layer. To select both layers hold the CTRL-key.

4.2.4 Select **Background correction** from the list tools.

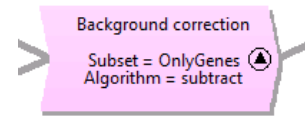
A description of this tool is displayed below the list of preprocessing tools.

4.2.5 Press <Add>. Double clicking on the tool in the list does the same.

4.2.6 Leave the settings in the next window unaltered and press <OK>.

4.2.7 The result of the background subtraction is stored in the **Target** layer.

4.2.8 Click on the arrow in the Background correction box. Two parameters are displayed.



4.2.9 To change the parameters, double click in the pink box.

4.2.10 In this example session, we will use the default settings (OnlyGenes and subtract). Press <Cancel>.

4.2.11 Repeat step 4.2.3 - 4.2.6 for the **Reference** and **ReferenceBackgr** layer.

The *Preprocessing* window should now look like Figure 4-5.

### • Variance stabilization

4.2.12 Select the **Target** layer on the right side of the Background correction box (see highlighted box in Figure 4-5).

4.2.13 Select **Variance stabilization** from the list of preprocessing tools and press <Add>. Double clicking on the name does the same.

4.2.14 The result of the variance stabilization is stored in the **Target** layer.

4.2.15 Select the **Reference** layer on the right side of the Background correction box.

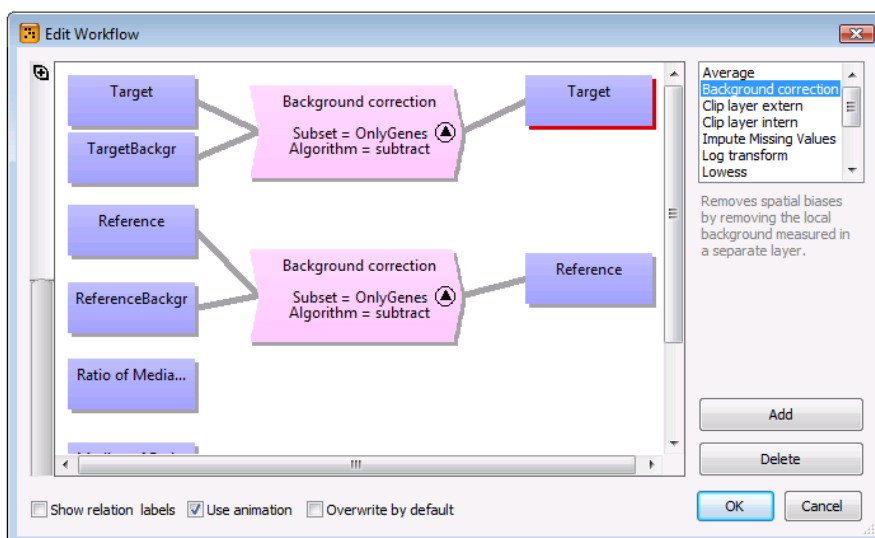


Figure 4-5. The *Preprocessing* window: Background correction.

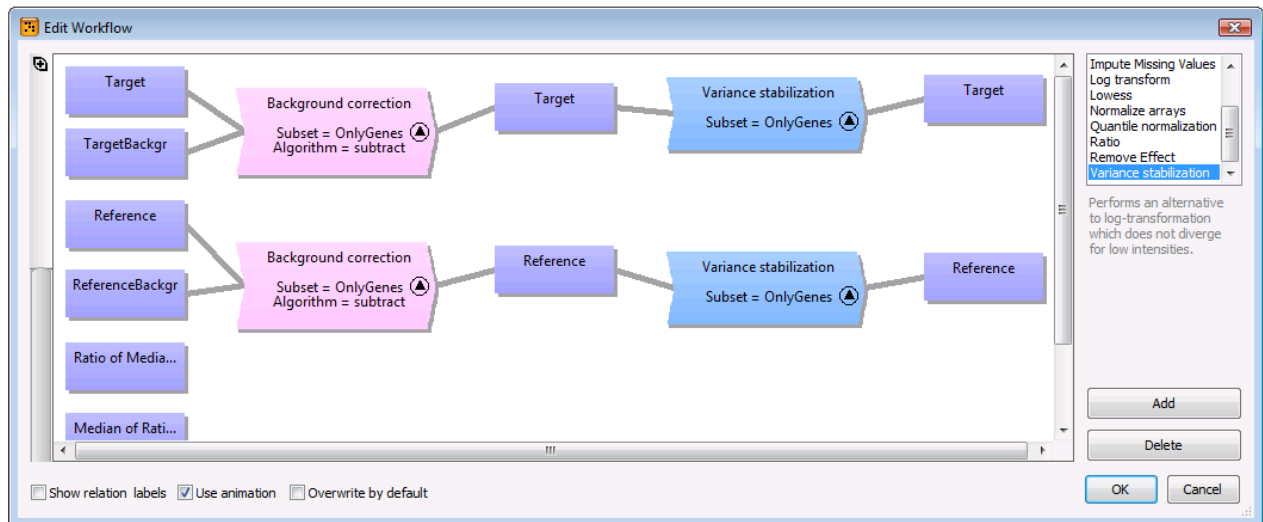


Figure 4-6. The *Preprocessing* window: Variance stabilization.

4.2.16 Select **Variance stabilization** from the list of preprocessing tools and press **<Add>**. Double clicking on the name does the same.

4.2.17 The result of the variance stabilization is stored in the **Background** layer.

The *Preprocessing* window should now look like Figure 4-6. Check if the parameter of the Variance stabilization is set to 'OnlyGenes'.

#### •Normalization

In order to compare the arrays with each other, the arrays must be normalized before we can draw statistically valid conclusions.

4.2.18 Select the **Target** layer on the right side of the Variance stabilization box.

4.2.19 Select **Normalize arrays** from the list of preprocessing tools and press **<Add>**. Double clicking on the name does the same.

4.2.20 The result is stored in the **Target** layer.

4.2.21 Do the same for the **Background** layer.

The *Preprocessing* window should now look like Figure 4-7.

#### •Lowess

In a next preprocessing step, we are going to use a Lowess normalization to compensate for intensity dependent effects caused by a difference in sensitivity between the two channels.

4.2.22 In the *Preprocessing* window, select the **Target** and **Reference** layer on the right side of the window. To select both layers hold the CTRL key.

4.2.23 Select **Lowess** from the list of preprocessing tools and press **<Add>**. Double clicking on the name does the same.

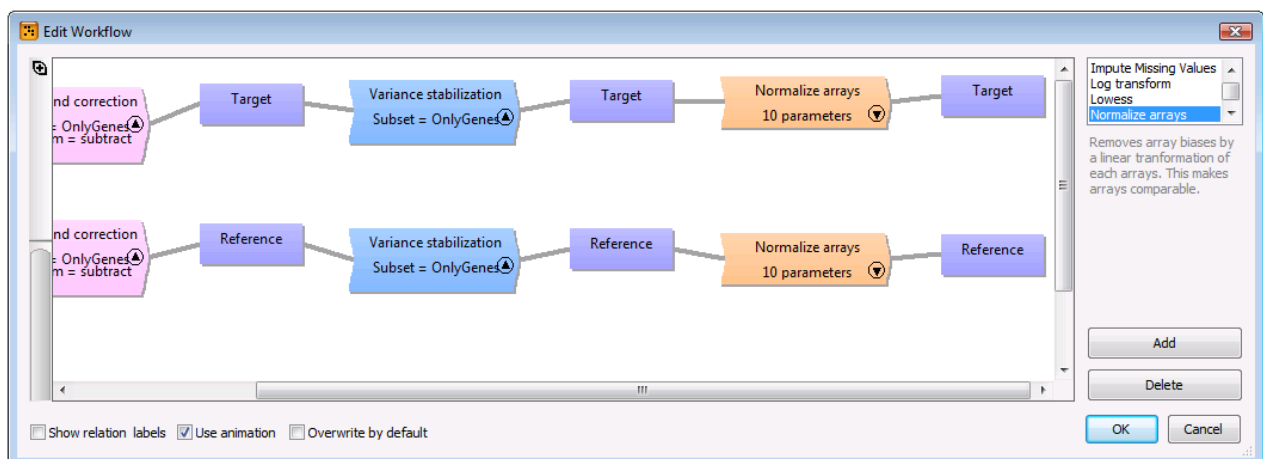


Figure 4-7. The *Preprocessing* window: Normalize arrays.

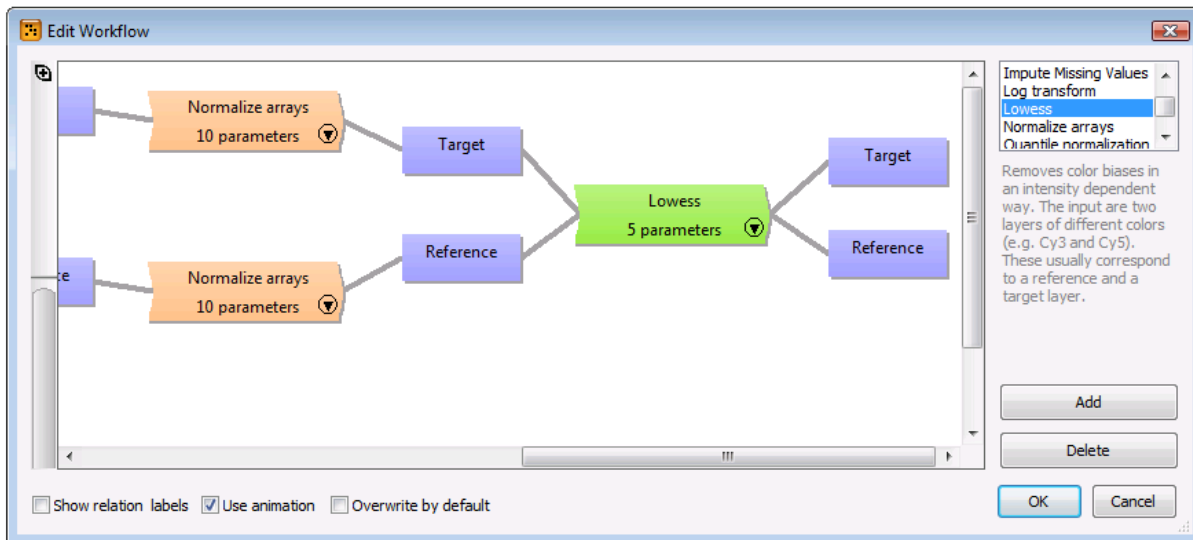


Figure 4-8. The *Preprocessing* window: Lowess.

The result of the Lowess normalization is stored in the **Target** and **Reference** layer.

#### • Ratio

In a last step we are going to calculate the ratio of the preprocessed **Target** and **Reference** layers.

4.2.24 Select the **Target** and **Reference** layers (on the right side of the Lowess boxes) in the *Preprocessing* window. To select both layers use the CTRL-key.

4.2.25 Select **<Add>**, press **<OK>** and enter a name for the layer containing the ratio values (e.g. **Ratio**). Press **<OK>**.

The *Preprocessing* window should now look like Figure 4-9.

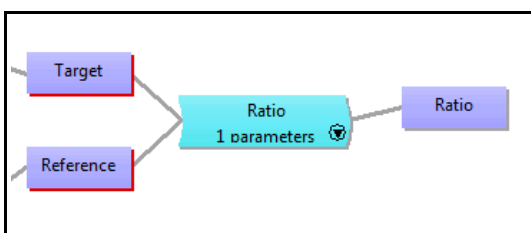


Figure 4-9. Store values in Ratio layer.

4.2.26 Press **<OK>**.

A *Calculation* dialog box pops (see Figure 4-10).

After calculating the preprocessing steps defined in the *Preprocessing* window, a new layer called **Ratio** is added to the list of layers in the *Layers* window (see Figure 4-11).

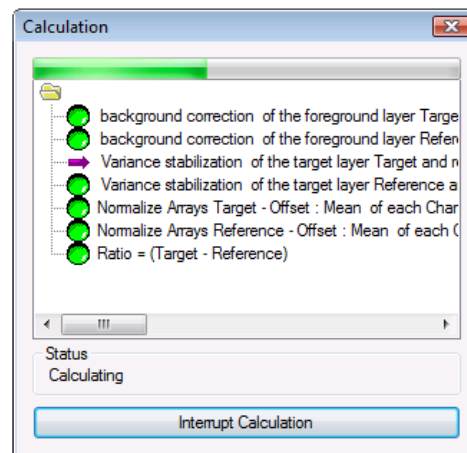


Figure 4-10. The *Calculation* dialog box.

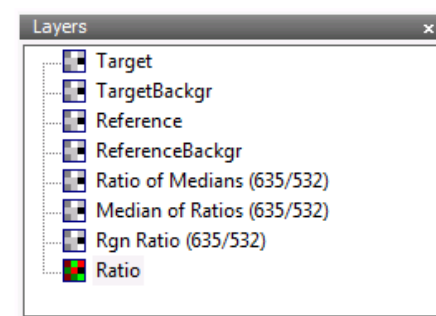


Figure 4-11. The *Layers* window with the new Ratio layer.

## 4.3 From spots to genes

In our session each gene has two or more spots on the microarray. Spots from the same gene have the same **ID** text field in the session.

If we want to work with the combined information of the two (or more) spots of each gene, we need to collapse the spots to genes. This can be done based on the content of the ID textfield in our session. The average is taken over the individual spots and these values are stored in a new root subset.

4.3.1 Select the row ID identifier from the list of row identifiers (if not already done).

4.3.2 Select *Subset > Collapse Aspect > By Field*.

4.3.3 Fill out the next dialog box as shown in Figure 4-12. Make sure the **OnlyGenes** subset is selected and press <OK>.

4.3.4 A new aspect called **Genes** is added to the session. The new scope **Genes x Arrays** is added to the *Subset* window.

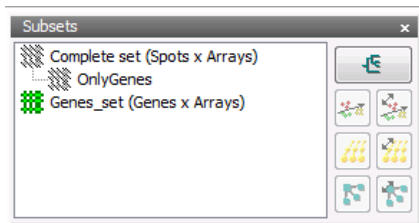


Figure 4-13. The *Subsets* panel.

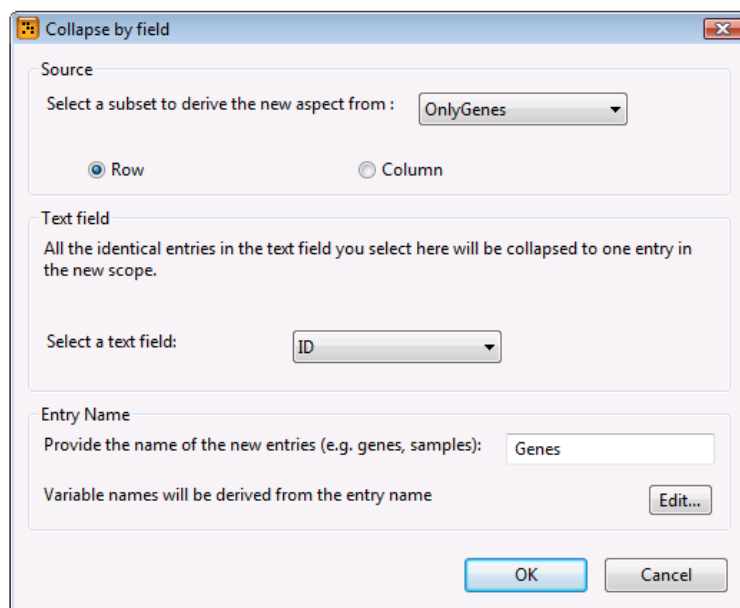


Figure 4-12. Create new aspect: Genes.

With the new scope selected, the averaged values are shown in the matrix panel.

4.3.5 Click on the row identifier tab (see Figure 4-14).

Only if the content in the row identifiers is the same for all spots of each gene, the row identifiers are transferred to the new aspect. Three row identifiers are transferred to the genes aspect (see Figure 4-14).

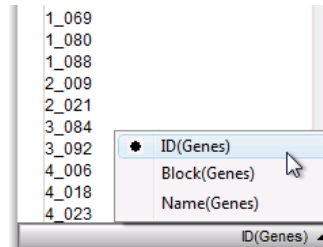



Figure 4-14. Row identifiers for the genes aspect.

## 4.4 Arrange columns

In a next step, we are going to arrange the columns based on the values of the information field 'Week'.

4.4.1 Make sure the 'Genes\_set (Genes x Arrays)' scope is selected in the *Subsets* panel.

4.4.2 Select *View > Arrange Column* or press the  button.

4.4.3 In the *Arrange Column* dialog box, select the field **Week** in the *By Field* tab and press <OK> (see Figure 4-15).

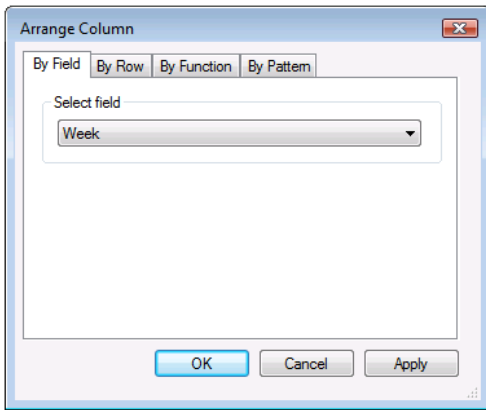


Figure 4-15. Arrange columns by field.

The arrays are now arranged in the *Column names* panel based on their Week values:

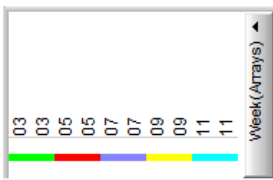


Figure 4-16. Arranged arrays.

## 4.5 Profile

Next, we are going to make a profile of the values of the **Dose** column information field.

4.5.1 Select **Dose** from the list of column identifiers.

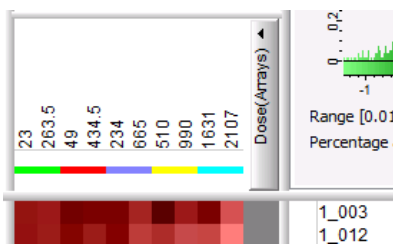


Figure 4-17. The Dose identifier.

4.5.2 The values are shown in the *Column names* panel (see Figure 4-17).

4.5.3 Select *Textfields > To Profile*.

4.5.4 In the next window, select the aspect **Arrays** and select **Dose** from the drop-down menu (see Figure 4-18). Press **<OK>**.

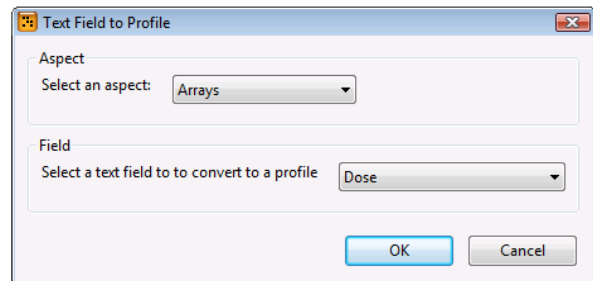


Figure 4-18. Choose text field to convert to profile .

4.5.5 Select the newly created profile in the *Info* panel on the right side of the window. Right click on the profile name and select **Store Profile**.

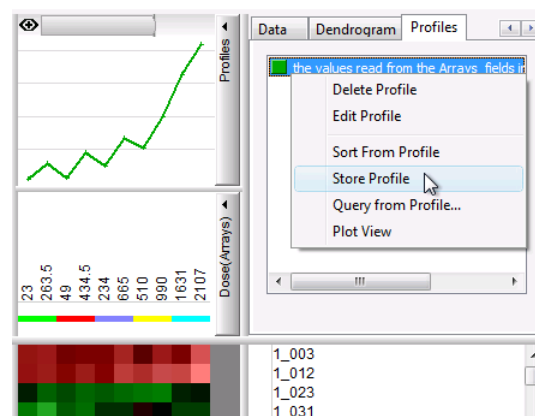


Figure 4-19. Store profile.

Name the profile **Dose** and press **<OK>**.

## 5. Statistics & Analysis


### 5.1 Statistic tests on the column groups

Often, the results from a statistical test are used to select genes or arrays that carry relevant information in the context of ongoing research. Usually, a threshold for the p-value obtained from the statistical test is used to make a selection. Statistical tests on populations are available in the *Statistics* wizard.

#### • ANOVA

First we are going to look for genes that are differentially expressed between the 'Week' groups. We are going to use an ANOVA test because there are more than 2 groups.

5.1.1 Make sure no entries are selected (press F4).

5.1.2 Select *Profiles > Statistics Wizard* or press .

5.1.3 Make sure the first option is selected in the *Orientation* panel, select **Genes\_set** in the *Subset* panel and press *<Next>* (see Figure 5-1).

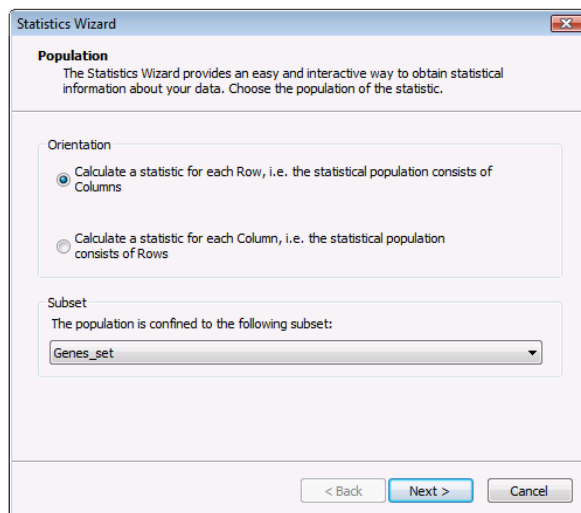


Figure 5-1. *Statistics wizard: Step 1.*

5.1.4 In the next step, open the *Independent test (multiple groups)* root element and select **ANOVA test** (see Figure 5-2). A short description of the statistic is provided on the right side of the window. Click *<Next>*.

5.1.5 In the next window, the parameters for the test need to be specified. Make sure that **Ratio** is selected in

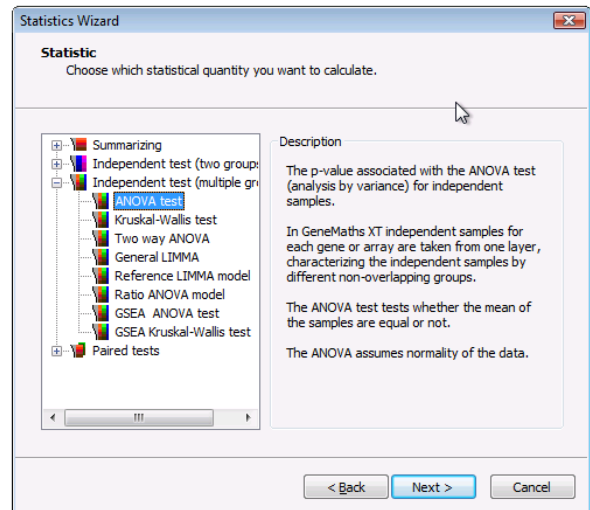


Figure 5-2. *Statistics wizard: Step 2.*

the *Layer* panel and **Week** in the *Groups* panel. Select **p-value** as output and click *<Next>*.

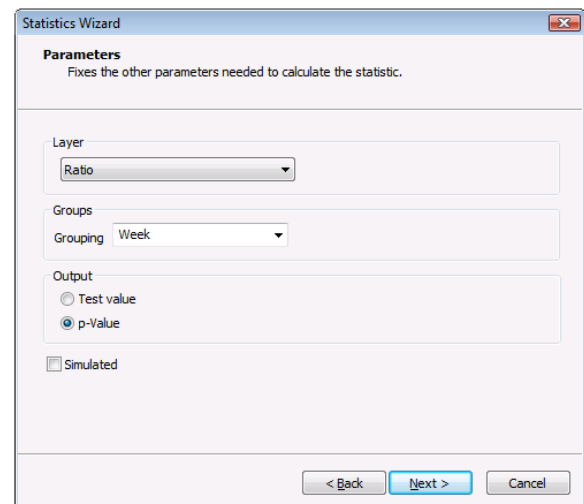


Figure 5-3. *Statistics wizard: Step 3.*

5.1.6 In the last window it is possible to specify a correction in case of multiple testing. For this exercise, select **None** and press *<Finish>*.

5.1.7 Click on the newly created profile in the *Profiles* tab. Right click on the profile name and select **Sort From Profile**.

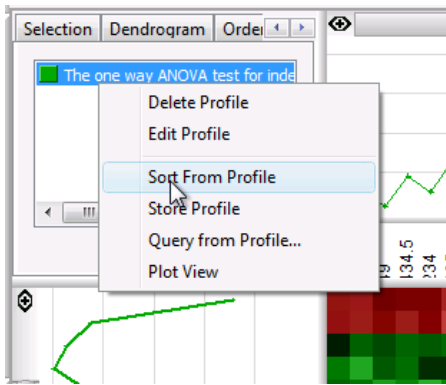


Figure 5-4. *Sort From Profile*

5.1.8 Right click in the *Profile* panel and select *Show as Numbers*.

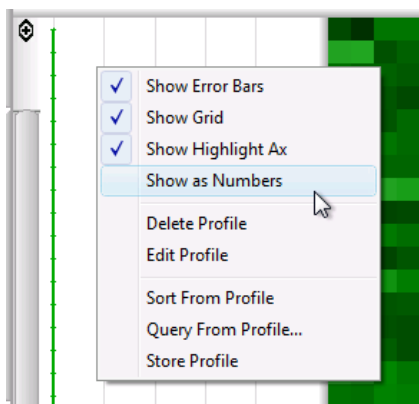


Figure 5-5. *Show as Numbers*.

In the *Profile* panel, the p-values for the genes are shown. These p-values give an indication if genes are significantly differentially expressed between the different 'Week' groups or not. The lower the p-values the more differentially expressed.

5.1.9 Right click in the *Profile* panel and select *Query From Profile*.

5.1.10 Set the threshold of the p-values to '< 0.01' and press <OK>.

All genes with a p-value smaller than 0.01 are selected (blue arrow in the *Row names* panel).

5.1.11 Select *Selection > Store Selection* and store the selection in a *New destination query* called **weekdiff**. Select **Genes\_set** and **Row** and press <OK>.

5.1.12 Unclear the selection by pressing F4.

## •WELCH T-TEST

Next, we are going to look for genes that are differentially expressed between the 'Batch' groups. We are going to use a Welch t-test. .

5.1.13 Select *Profiles > Statistics Wizard* or press



5.1.14 Make sure the first option is selected in the *Orientation* panel, select **Genes\_set** in the *Subset* panel and press <Next> (see Figure 5-6).

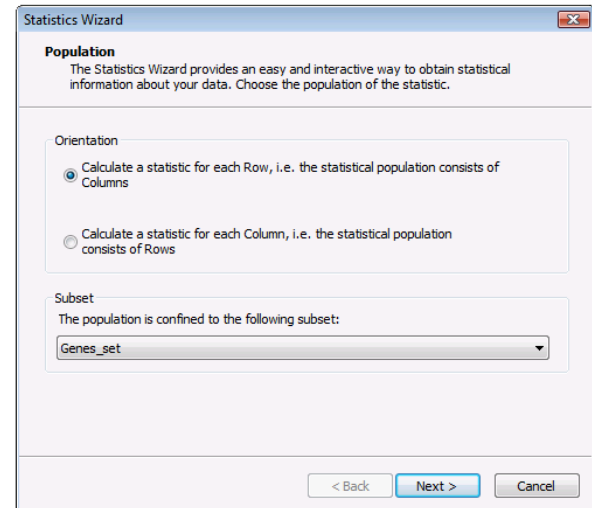


Figure 5-6. *Statistics wizard: Step 1*.

5.1.15 Select *Independent Welch t-test* (under 'Independent test (two groups)') from the list and click <Next>.

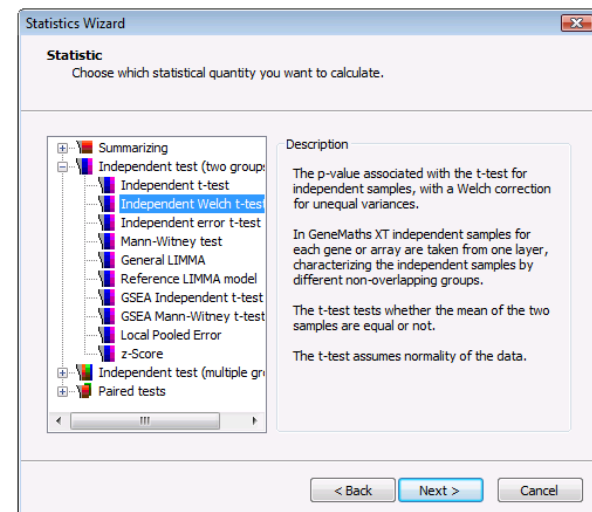


Figure 5-7. *Statistics wizard: Step 2*.

5.1.16 In the next window, make sure that **Ratio** is selected in the *Layer* panel and **Batch** in the *Groups* panel. Select **p-value** as output and click <Next>.

5.1.17 In the last window, select **None** and press <Finish>.

5.1.18 Click on the newly created profile in the *Profiles* tab. Right click on the profile name and select *Sort From Profile*.

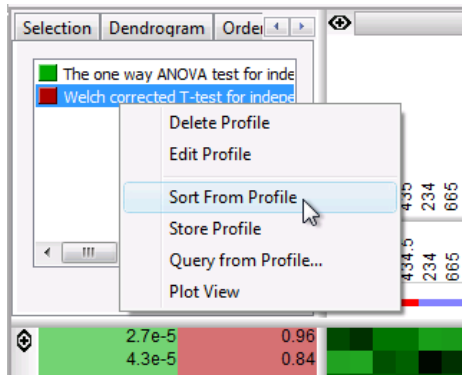


Figure 5-8. Sort From Profile.

5.1.19 Right click in the *Profile* panel and select *Show as Numbers* (if not already shown).

In the *Profile* panel, the p-values for the genes are shown. These p-values give an indication if genes are significantly differentially expressed between the different 'Batch' groups or not. The lower the p-values the more differentially expressed. In this session, most genes have a high p-value, indicating that the batch effect is not that strong.

## 5.2 Remove effect

In a next step, we are going to remove the small batch effect.

5.2.1 Select *Layer > Normalization > Remove Effect*.

5.2.2 In the next dialog box, fill out the settings as shown in Figure 5-9 and press <OK>.

With these settings, the batch grouping effect is removed.

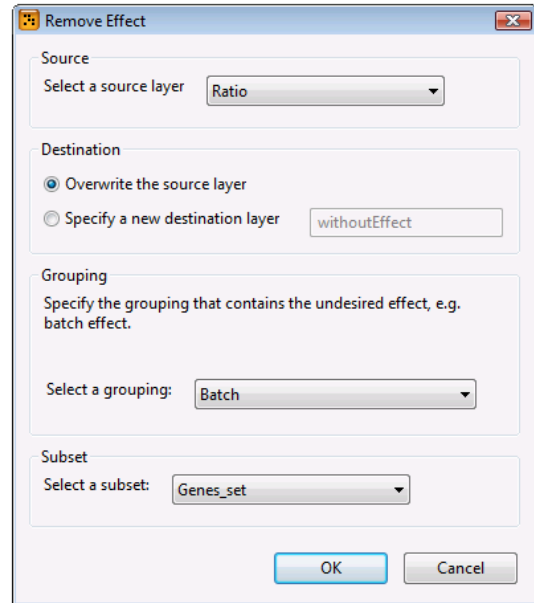


Figure 5-9. Remove the Batch effect.

## 5.3 Pattern matching

Next, we are going to look for the row entries whose expression values follow the **Dose** pattern as closely as possible.

5.3.1 The pattern matching tool is opened from the Main view using *Profiles > Pattern Matching* (see Figure 5-10).

5.3.2 Select the **Dose** profile from the list under *Stored profiles* and press the <Rank rows> button (see Figure 5-10).

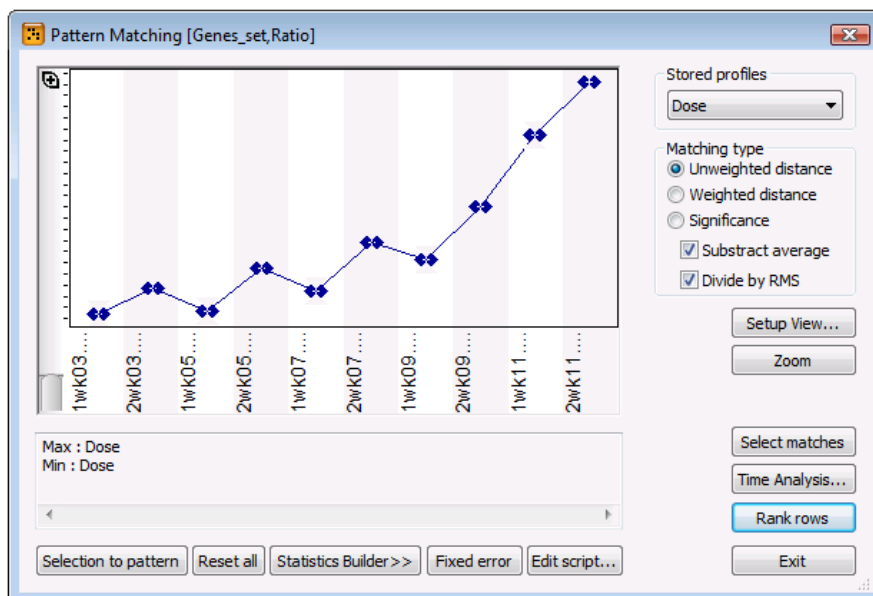


Figure 5-10. The *Pattern Matching* window.

The distance of each row (= each gene) to the **Dose** pattern is calculated. The rows are ranked according to increasing distance to the pattern. The lower the value, the better the match.

5.3.3 Right click in on the newly created profile and select *Query from Profile*.

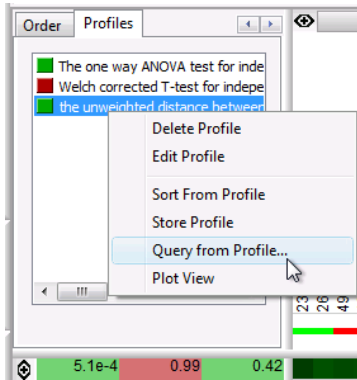


Figure 5-11. Query from Profile.

5.3.4 Set the threshold of the values to '< 1' and press <OK>.

All row entries with a value smaller than 1 are selected (blue arrow in the *Row names* panel).

5.3.5 Select *Selection > Store Selection* and store the selection in a *New destination query* called **likedose**. Make sure **Genes\_set** and **Row** are selected. Press <OK>.

5.3.6 Clear the selection by pressing F4.

## 5.4 Venn diagram

With the Venn diagram option in GeneMaths XT we will plot the following two stored selections:

- **weekdiff**: differentially expressed genes between the week groups.
- **likedose**: genes that resemble the Dose pattern as close as possible.

5.4.1 Select *Selection > Venn Diagram*. Select the two stored queries from the drop down menus and press <OK>.

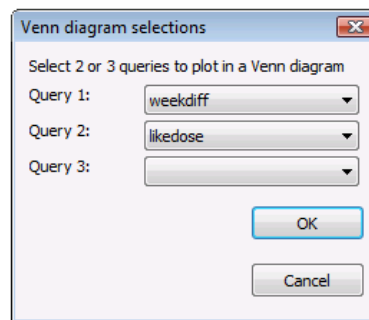


Figure 5-12. Venn diagram selections.

The Venn diagram is shown in a new window. 26 genes are shared amongst the two tests (see Figure 5-13).

5.4.2 Click on the number of shared genes in the Venn diagram. Select <Yes> to confirm that you want to select all the genes that are shared amongst the two tests.

5.4.3 Close the Venn diagram.

The shared genes are selected in the main window.

5.4.4 Press F4 to unselect all genes.

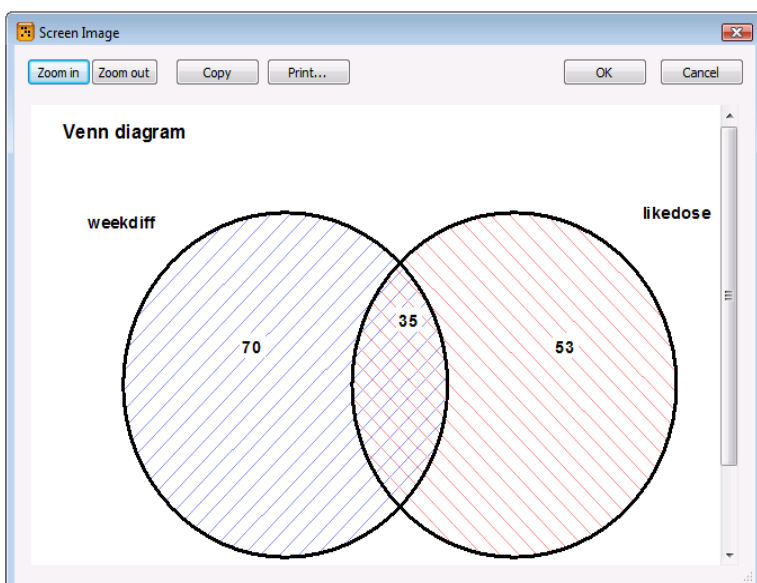


Figure 5-13. Venn diagram.

## 5.5 Row grouping

In chapter 3 we have created two column groupings. In this section we will create a row grouping. The groups will be based on the 'Name' information field.

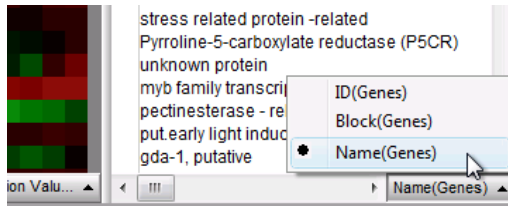


Figure 5-14. The 'Name' identifier.

5.5.1 Make sure the **Ratio** layer and the **Genes\_set** scope is selected in the main window.

5.5.2 Select the **Name** row identifier (see Figure 5-14).

5.5.3 Select **Groups > Edit Row Groups** and click on **<Create New Grouping>**.

5.5.4 In the next window, select **Name** from the drop-down menu and press **<OK>**.

5.5.5 In the next window, check **Minimum number of entries** and enter 2. Unselect the other two checkboxes and press **<OK>** (see Figure 5-15).

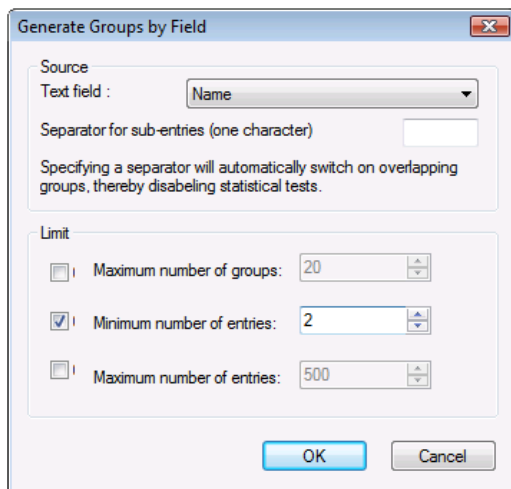


Figure 5-15. Group settings.

5.5.6 The groups based on the settings are shown in the next window. Press **<Exit>**.

5.5.7 Drag the **Group** window next to the **Layers** window by clicking on the caption of the **Group** window and holding the mouse button.

5.5.8 Select **Name** from the drop down menu in the **Group** panel.

The groups of the Name grouping are listed below the drop-down menu (see Figure 5-16). Group members can be selected from within this panel.

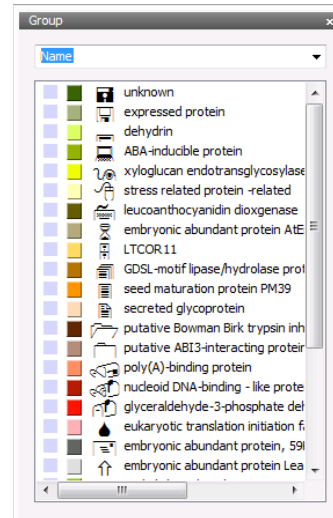


Figure 5-16. Part of the **Group** window with the groups based on the 'Name' information field.

5.5.9 Select the first two groups (**unknown** and **expressed protein**) in the **Groups** window using the CTRL-button. All entries belonging to these two groups are selected in the **Row names** panel.

5.5.10 Select **Selection > Invert Row Selection**.

The selection is inverted.

5.5.11 Select **Subset > Selection to Subset**. Give a name to the subset e.g. **OnlyKnownGenes**, make sure **Child of the current subset** is selected and press **<OK>**.

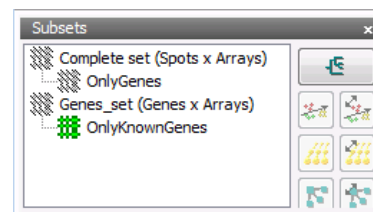


Figure 5-17. The **Subsets** window.

The new subset is listed in the **Subsets** window as a child of the **Genes\_set**.

5.5.12 Press F4 to clear the current selection.

## 5.6 Hierarchical clustering

Hierarchical clustering is a popular technique used for the grouping of entries. GeneMaths XT offers a wide range of similarity coefficients and clustering methods, both for clustering rows and columns.

5.6.1 Select *Analysis > Cluster Analysis*. In the window that pops up, select **Rows**, the **Ratio** layer, and **OnlyKnownGenes** in the upper panel of the window. Leave the lower panel unaltered (see Figure 5-18).

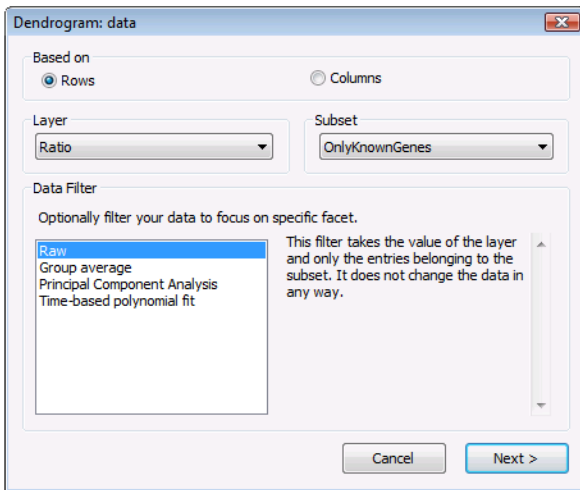


Figure 5-18. Cluster analysis: step 1.

5.6.2 Choose the **Pearson correlation** coefficient in the next window (see Figure 5-19). Press **<Next>**.

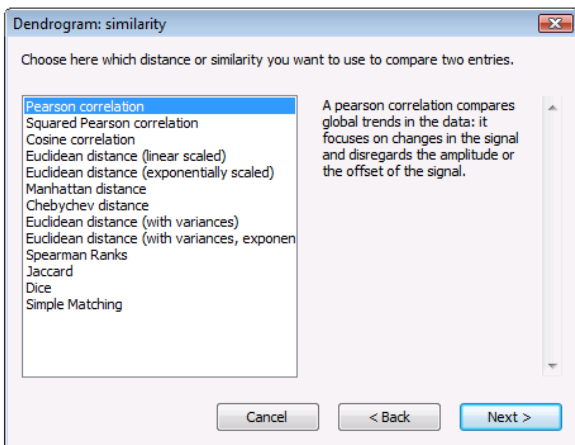


Figure 5-19. Cluster analysis: step 2.

5.6.3 In the next window, choose **UPGMA** as clustering method (see Figure 5-20). Press **<OK>**.

The dendrogram based on the rows is shown in the main window of GeneMaths XT. To get a better overview, use the vertical zoom slider.

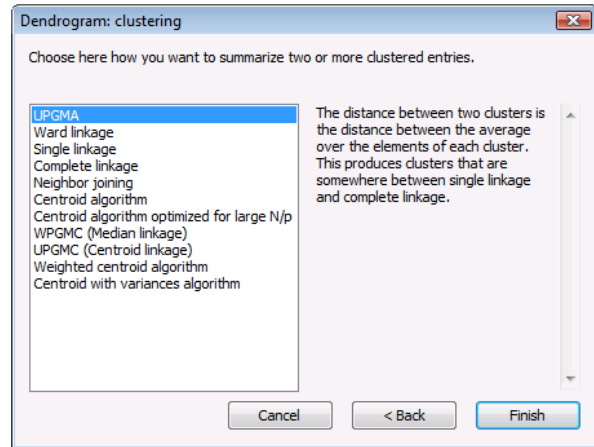


Figure 5-20. Cluster analysis: step 3.

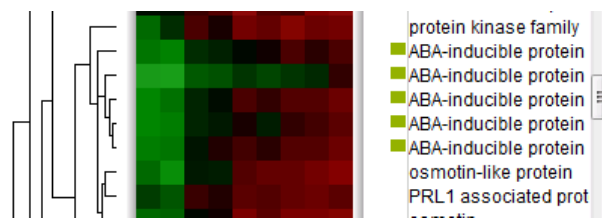


Figure 5-21. Clustering the row entries.

5.6.4 Select the **Name** row identifier.

The dendrogram clearly clusters genes with the same 'Name' together (see Figure 5-21).


Next, we are going to cluster the columns.

5.6.5 Select *Analysis > Cluster Analysis*. In the window that pops up, select **Columns**, the **Ratio** layer, and **OnlyKnownGenes** in the upper panel of the window. Leave the lower panel unaltered.

5.6.6 Choose the **Pearson correlation** coefficient and the **UPGMA** cluster method in the following steps.

The dendrogram based on the columns is shown in the main window of GeneMaths XT.

5.6.7 Select the column identifier tab and select the **Week** column identifier from the list.

5.6.8 To see the colors based on the Week grouping (if not already shown), click on the arrow next to the  button and select **Week** from the drop-down menu.

The column groups based on the week number are clearly separated from each other (see Figure 5-22).

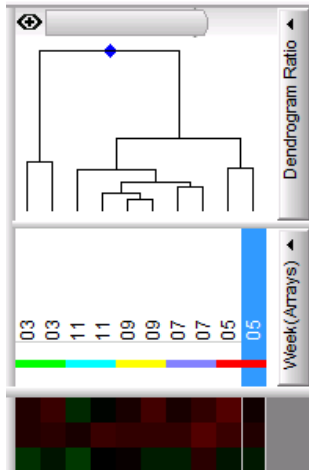



Figure 5-22. Clustering the columns.

## 5.7 Dimensioning techniques

Next, we are going to perform a principal components analysis (PCA).

5.7.1 Make sure the **Week** column identifier is selected.

5.7.2 Click on the arrow next to the  button and select **Week** from the drop-down menu.

5.7.3 Select **Analysis > Principal Components Analysis**

or press the  button.

5.7.4 In the next window, select **Columns**, the **Ratio** layer and the **OnlyKnownGenes** subset in the upper panel. Leave the other settings unaltered. Press **<OK>**.

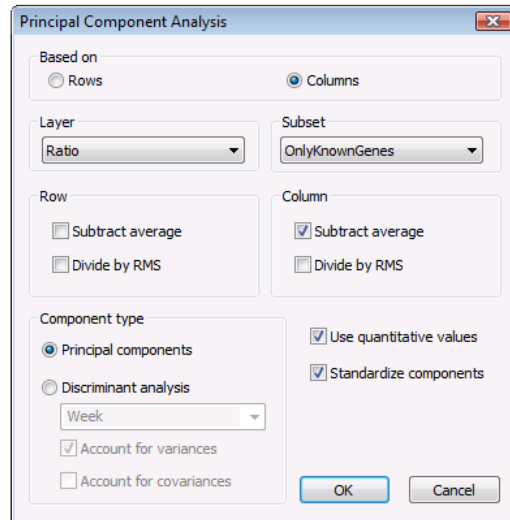


Figure 5-23. Dimensioning settings.

The result of this analysis is shown in Figure 5-24. The left part of this PCA view contains a number of information panels. The middle panel shows the projection of the columns on the principal components, with colors corresponding to the current column grouping (in this case the 'Week' grouping). The projection of the genes

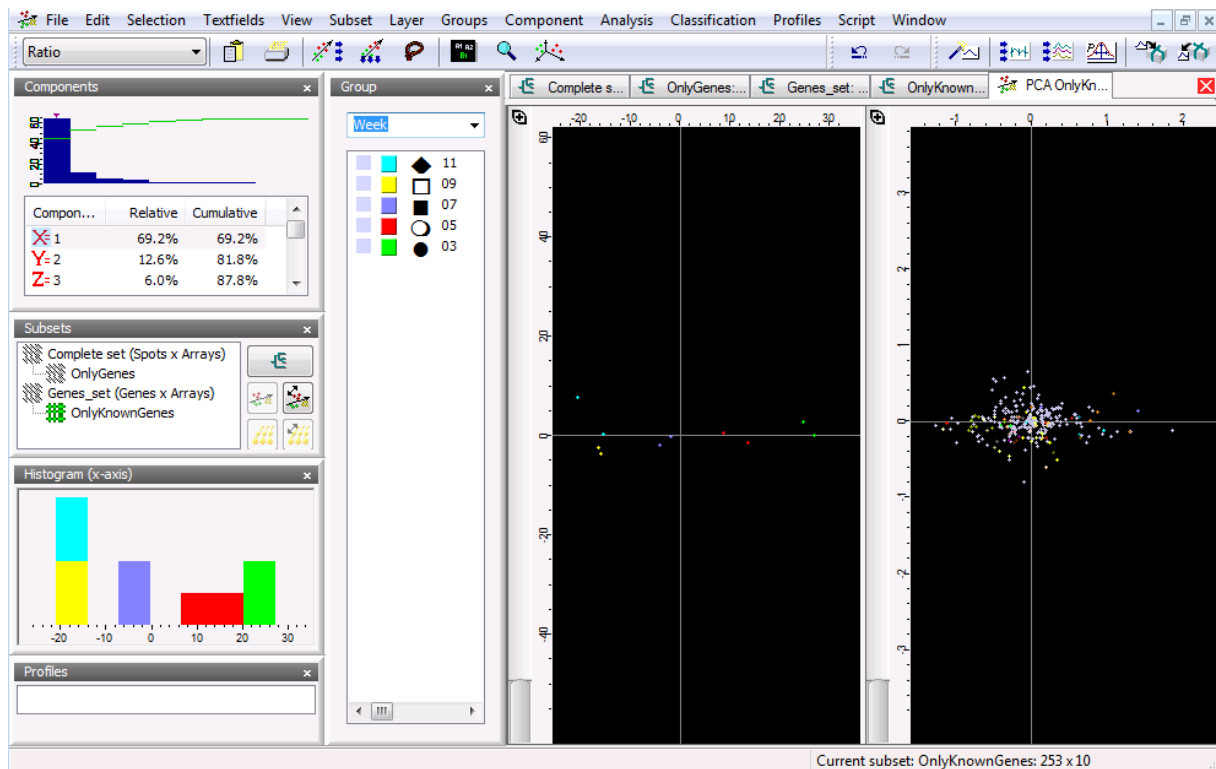


Figure 5-24. The PCA view.

on the principal components is shown in the right panel, with colors corresponding to the current row grouping (the 'Name' grouping).

From the view in Figure 5-24, we can clearly see that the members of the different groups of the 'Week' grouping are grouped together, indicating that the week group have different effects on the row entries (=genes). Those column entries occurring left on the X-axis are positive for the left row entries and negative for the right row entries. Column entries occurring right on the X-axis are positive for the right row entries and negative for the left row entries.

5.7.5 Select a few entries by holding the CTRL-button.

The selected entries are surrounded by a blue square.

5.7.6 Go back to the *Main* view.

The entries selected in the *PCA* view are also selected in the *Main* view.

## 5.8 Partitioning

Calculating a partitioning is a non-hierarchical method for grouping of entries. To obtain a partitioning that makes sense, one should have a rough idea about the number of groups that are expected in the data set.

5.8.1 Open a partitioning with *Analysis > Partitioning*.

5.8.2 Select **Rows**, the **Ratio** layer and the **OnlyKnownGenes** subset in the next window. Press *<Next>* twice.

5.8.3 Select the cell.

5.8.4 Select *Partitioning > Split Cell*, set the number of partitions to 5 and press *<OK>*.

The result of the partitioning is shown in Figure 5-25.

5.8.5 The presentation of the cells can be changed with *View > Group Pie Chart* into charts that indicate the group membership of the profiles in the cells.

5.8.6 Right click on the cell selected in Figure 5-27 and select *Cell to Statistics Report*.

A detailed report pops up. In the report of the cell selected in Figure 5-27, the dehydrin group has a low p-value. This means that the odds of having that many entries belonging to the dehydrin group in this cell is much more than just by chance.

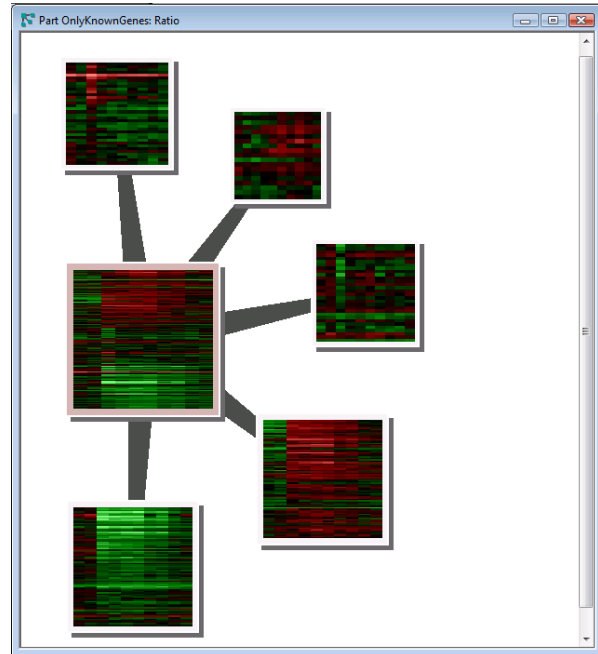


Figure 5-25. Partitioning presented with expression values.

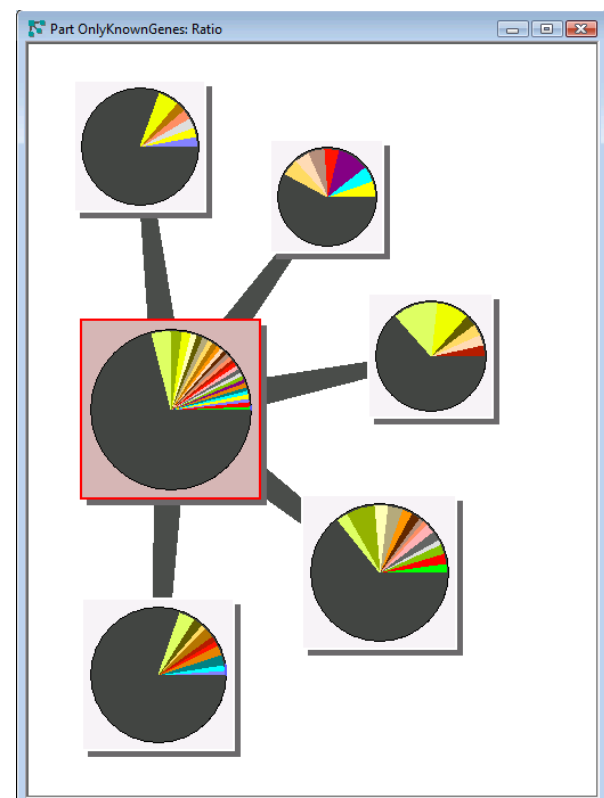


Figure 5-26. Groups from partitioning cells.

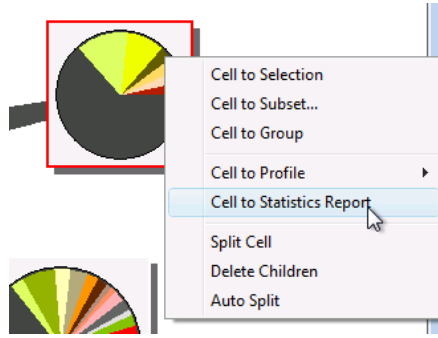


Figure 5-27. Cell to Statistics Report.

### 5.9 Self-Organizing map

Another grouping technique is the Self-Organizing Map (SOM). A SOM is a neural network that classifies entries in a two-dimensional map according to their likeliness. Just like with the partitioning tool, one should have an

idea of the number of expected groups in order to obtain a good result.

5.9.1 A SOM is calculated with *Analysis > Self-Organizing Map*.

5.9.2 Select **Rows**, the **Ratio** layer and the **OnlyKnownGenes** subset in the next window. Press **<Next>** twice.

5.9.3 Calculate a SOM with 5 cells in the X dimension and 3 cells in the Y dimension.

The SOM is calculated and shown (see Figure 5-28). Areas of high similarity have a dark shading.

5.9.4 To show the group membership of the cell entries select *View > Group Pie Chart* (see Figure 5-29).

5.9.5 Right click on the cell bordered with the yellow rectangle in Figure 5-29 and select *Cell to Statistics report*.

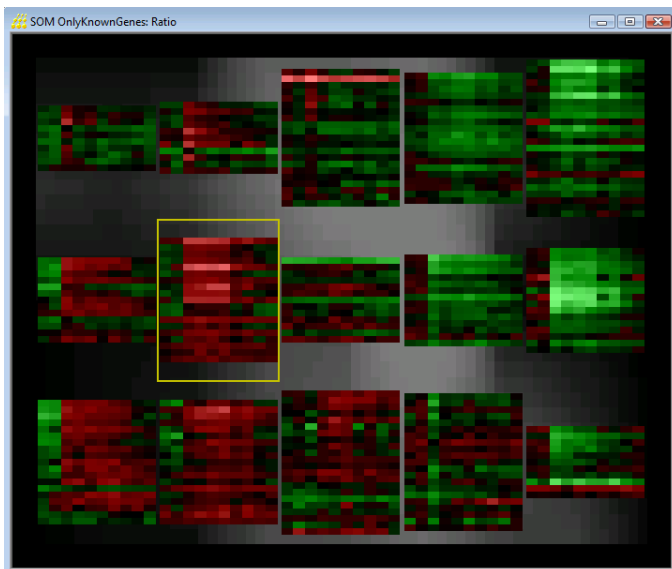


Figure 5-28. A SOM with 5 cells in the X dimension and 3 cells in the Y dimension.

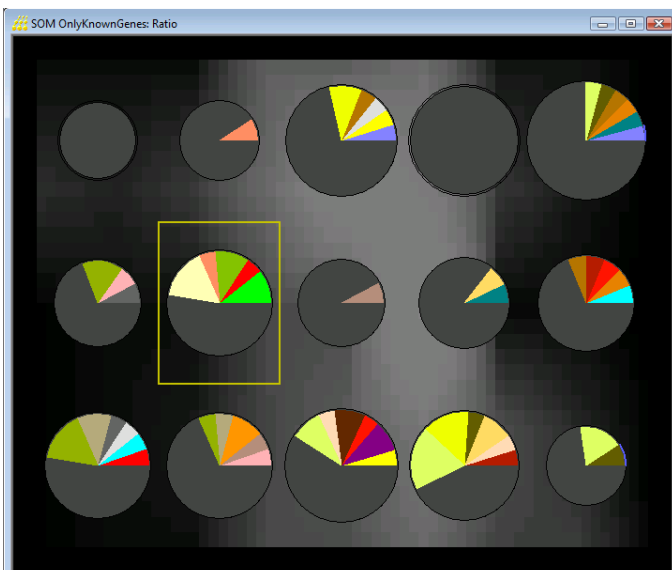


Figure 5-29. Group membership view.

In the report of the cell, the 'stress-related protein, '60S ribosomal protein L15', and early light-induced protein' groups have a low p-value. This means that the odds of

having that many entries belonging to these groups in this cell is much more than just by chance.