GeneMaths XT

Time course tutorial

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

NOTES

SUPPORT 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

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

URL: 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

1. Import 5

1.1 Downloading the data 5

1.2 Importing the data in GeneMaths XT. 5

2. Annotation 9

2.1 Row annotation 9

2.2 Column annotation 10

3. Groupings. 13

3.1 Row groups 13

3.2 Column groups 15

4. Preprocessing. 17

4.1 The preprocessing diagram 17

4.2 Log transformation 17

4.3 Normalize arrays. 17

4.4 Normalize genes 18

5. Statistics & Analysis. 19

5.1 Statistical analysis on the ‘timepoint’ grouping 19

5.2 Statistical analysis: control versus heat resistant 22

5.3 Cluster analysis 23

5.4 Creating a self organizing map (SOM) and a partitioning. 25

1. Import

DNA microarrays provide a wealth of information about biology at the level of gene expression and is a powerful tool to identify genes and pathways involved in various processes. Sørensen et al.(1) analysed the full heat stress reponse in *Drosophila melanogaster* females, using whole genome gene expression arrays (Affymetrix Inc, Santa Clara, CA, USA). The study focuses on up – as well as downregulation of genes from just before and at 8 time points after an application of short heat hardening. Four replicates are used: 2 control lines and 2 heat selected lines. This example dataset will be used in order to explain the workflow of GeneMaths XT. This dataset is publicly available on the GEO website.

1.1 Downloading the data

1.1.1 Go to the GEO homepage: http://www.ncbi.nlm.nih.gov/geo, click in the box next to 'Query > GEO accession' and type **GDS2272**.

1.1.2 Press **<*Go*>**.

1.1.3 Select **GSE5147** in the *GDS Summary* panel next to Series.

1.1.4 Scroll down the next page and select ***SOFT formatted family file(s)***.

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Figure 1-1. Download information.

1.1.5 On the next page select **GSE5147_family.soft.gz**.

1.1.6 Select **<*Save*>** and navigate to the path on your computer**.**

1.1.7 Press **<*Save*>** to save the file in the selected folder.

(1) Sørensen et al., Full genome gene expression analysis of the heat stress response in *Drosophila melanogaster*, *Cell Stess & Chaperones,* 10, 4, (2005).

1.2 Importing the data in GeneMaths XT

1.2.1 Start GeneMaths XT by double clicking on the icon on the desktop or from the task bar with ***Start > Programs > Applied Maths > GeneMaths XT.***

1.2.2 Click **<*Next*>** in the welcome screen to begin the import of the data. If the welcome screen does not appear, choose ***File > Import Wizard*** in the *GeneMaths XT Main* window. The *Import* Wizard window pops up (see Figure 1-2).

1.2.3 Select the fourth option ***Import from other sources*** and hit **<*Next*>**.

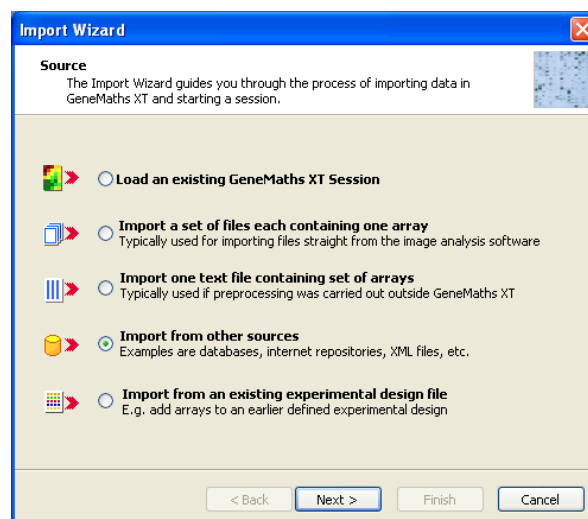


Figure 1-2. *Import* wizard: select data source.

1.2.4 Select ***GEO's SOFT family*** in the format list. A short description of the format is shown in the right panel (see Figure 1-3).

1.2.5 Click **<*Next*>**.

1.2.6 In the next window, browse for GSE5147_family.soft in the *File* panel and enable ***Time Course Experiment*** (see Figure 1-4). Click **<*Next*>**.

1.2.7 In the next window specify the name of the processed file e.g. GSE5147.xps and press the **<*Save*>** button.

1.2.8 The *Calculation* dialog box pops up. The status of the import of the data is shown at the bottom of the dialog box.

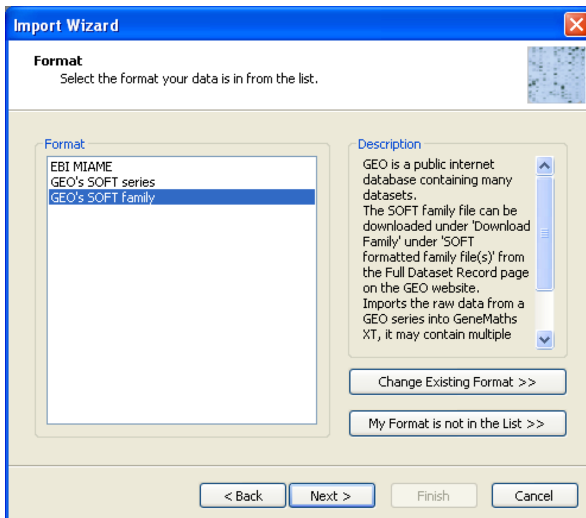


Figure 1-3. ***Import*** **wizard: select format.**

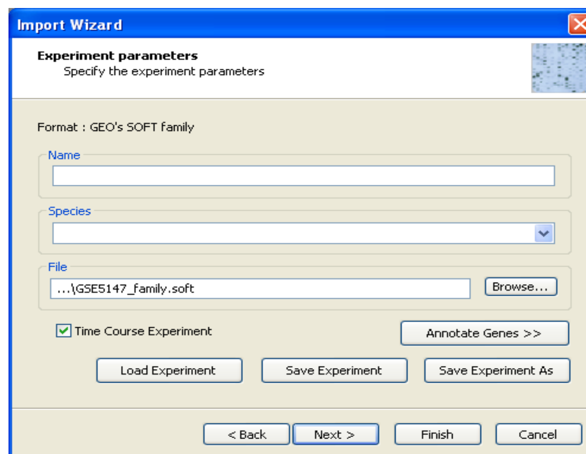


Figure 1-4. ***Start*** **Wizard: input file.**

1.2.9 A new window pops up. Leave the settings unaltered and press **<*OK*>**.

1.2.10 In the next window, select **ID-REF** press "**>**" and press **<*Next*>.**

1.2.11 In the next window, select **Value** and **No error** and press **">".**

GeneMaths XT will open a session with one layer called **VALUE** (see Figure 1-6).

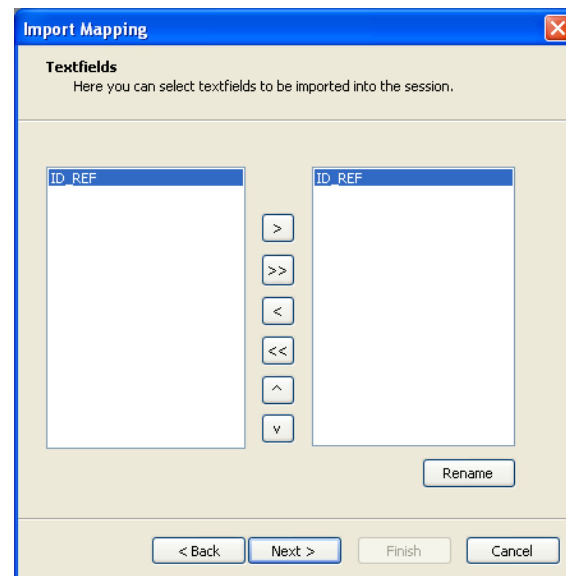


Figure 1-5. ***Import*** **wizard: mapping**

The session contains five column identifiers (see Figure 1-7) and sixteen row identifiers (see Figure 1-8).

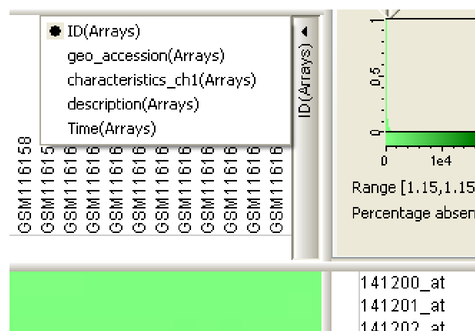


Figure 1-7. Five column identifiers.

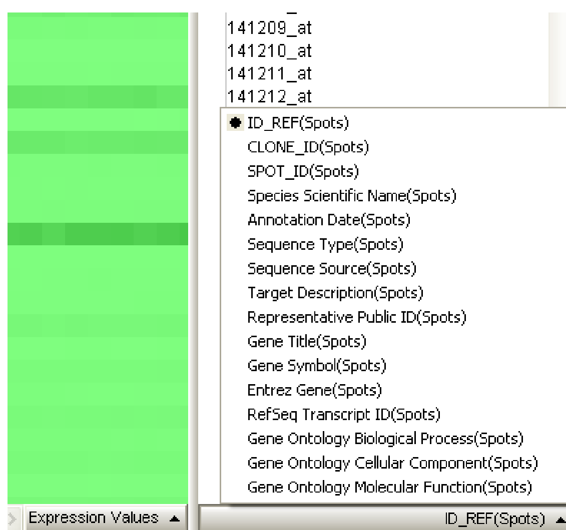
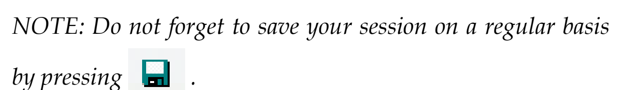


Figure 1-8. The row identifiers.

NOTE: Do not forget to save your session on a regular basis by pressing  .

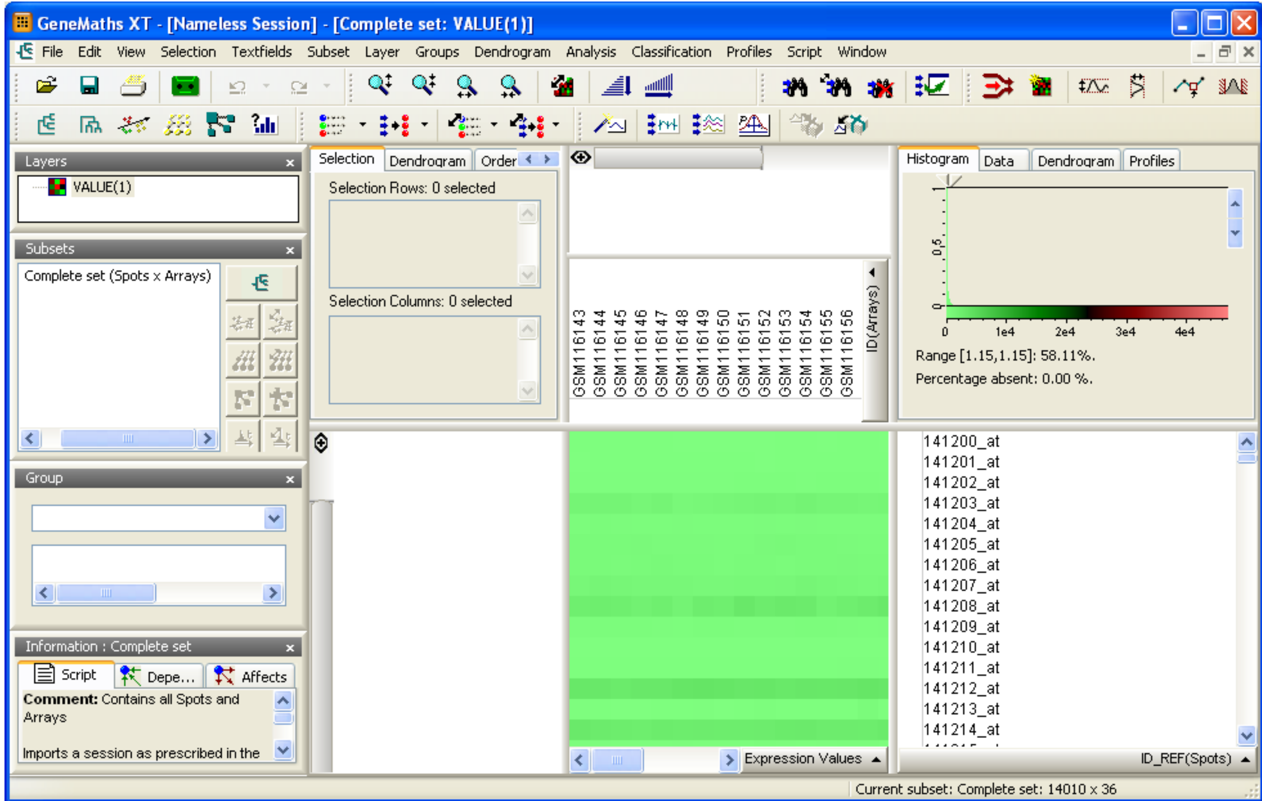


Figure 1-6. The *Main* window of GeneMaths XT after import of the data.

2. Annotation

2.1 Row annotation

2.1.1 Select ***Textfields > Annotations > Affymetrix*** in the *Main* window of GeneMaths XT and select **DrosGenome1** from the list. Leave ***GeneMaths GAF file*** enabled and make sure that GeneMaths XT will check for updates on the website (see Figure 2-1).

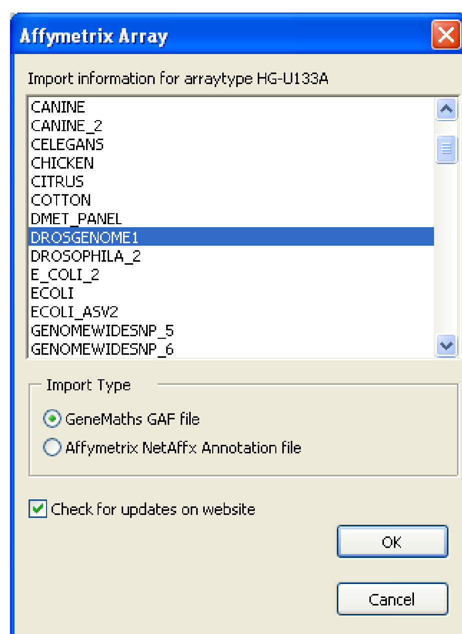


Figure 2-1. Select Affymetrix Arraytype.

2.1.2 Press ***<OK>***.

The processing may take a while depending on the speed of your internet connection.

2.1.3 If you are using the Affymetrix annotations for the first time you will need to subscribe yourself on the Affymetrix website (for free). Enter your credentials in the dialog box that pops up and press ***<OK>***.

2.1.4 In the next window, use **Probe Set ID** as a link for **ID_REF** and select **GO ID** (see Figure 2-2). Press ***<OK>***.

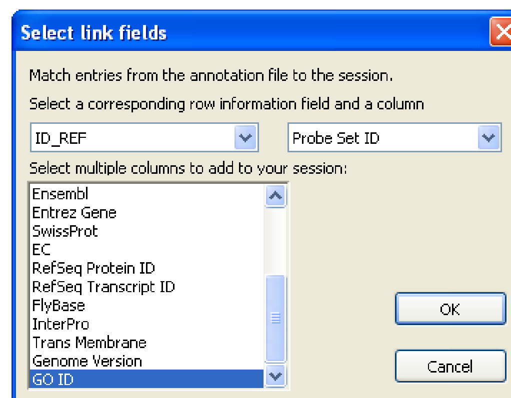


Figure 2-2. Adding identifiers to the gene identifier list.

2.1.5 Click on the bottom right tab in the *Main* window of GeneMaths XT (see Figure 2-3). The GO ID identifiers is added to the list of row identifiers.

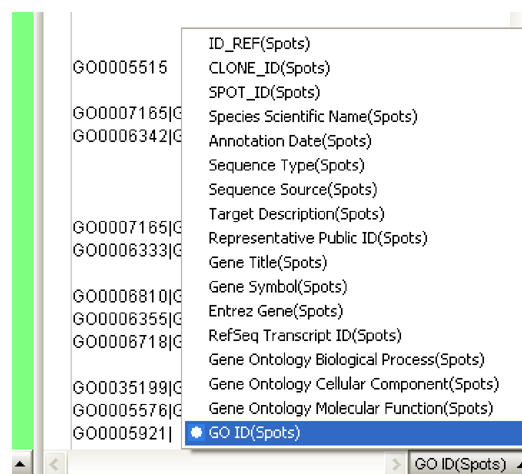


Figure 2-3. New information field added to the session.

In the next step we are going to add the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways to our genes.

2.1.6 Select ***Textfields > Annotations > KEGG***.

2.1.7 Select **Drosophila melanogaster (fruitfly)** as species and enable ***Download most recent version*** (Figure 2-4).

2.1.8 Press ***<OK>***.

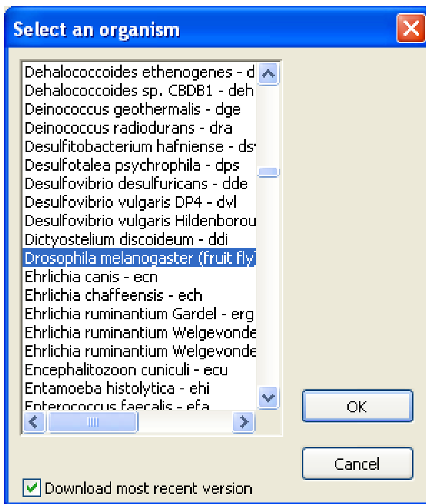


Figure 2-4. Select an organism.

2.1.9 Set the **Gene Symbol** information field as link (Figure 2-5) and press ***<OK>***.

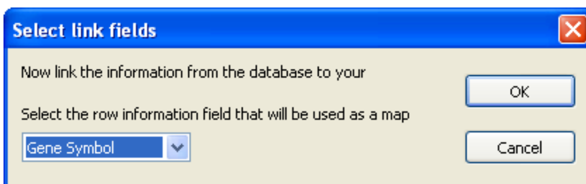


Figure 2-5. Linking the information.

KEGG Pathways is added to the list of Row identifiers (see Figure 2-6).

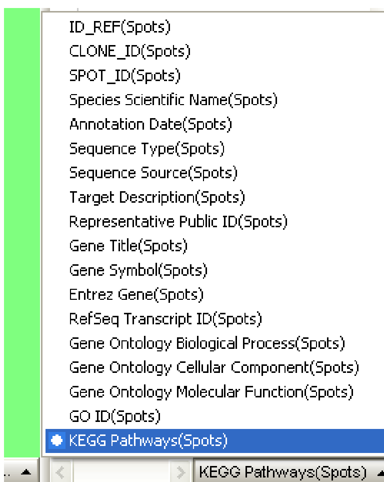


Figure 2-6. KEGG Pathways added to the list of Row identifiers.

2.1.10 Select **KEGG Pathways** from the list of Row Identifiers. The KEGG Pathways (if available) are shown in the *Row names* panel. One entry can contain multiple KEGG Pathways, separated by a " | ".

2.2 Column annotation

Five column identifiers are present after import of the data. In a next step, we are going to split the information present in the **characteristics_ch1** information field.

2.2.1 Select ***Textfields > Split.*** Fill out the dialog box as shown in and press ***<OK>***.

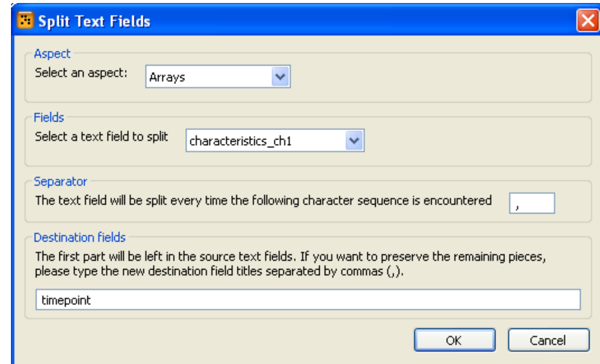


Figure 2-7. Split textfields.

One new column identifiers called **timepoint** is added to the list of column identifiers.

2.2.2 Select the **timepoint** identifiers from the list. The information in the *Column names* panel is updated (see Figure 2-8).

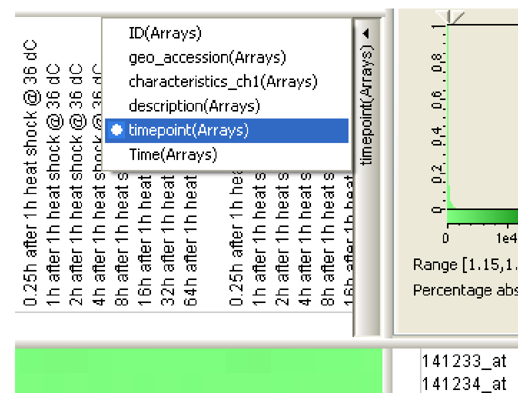


Figure 2-8. The timepoint identifier.

2.2.3 Select ***Script > Run Script*** and select the script **readTime.GXS**. Press ***<Open>***.

The script readTime.GXS reads the time as a GeneMaths XT variable.

2.2.4 A dialog box pops up, asking you to specify the array information field and the time format (see Figure 2-9). Select **timepoint** and **hours** and press ***<OK>***.

2.2.5 In the *Main* window of GeneMaths XT, select the **Time** column information field. The values are filled in correctly after running the script (see Figure 2-10).

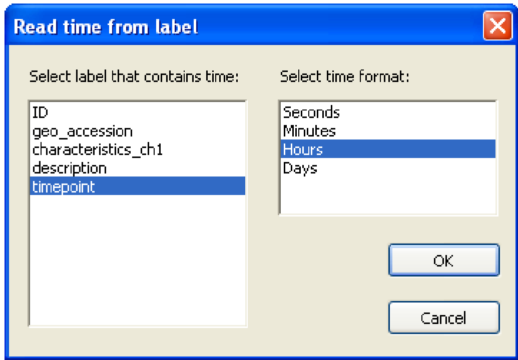


Figure 2-9. Select the column information field and the time format.

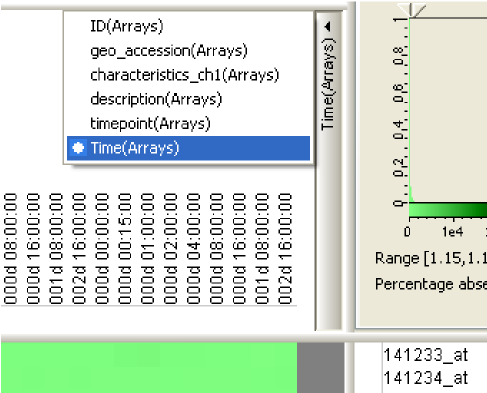


Figure 2-10. Values for the 'Time' column information field.

3. Groupings

With the statistics we want to perform later on in mind, we need to define groupings, each containing a set of particular groups. The groupings will then later be the input for the statistical tools and visualizations.

3.1 Row groups

In the first step, we are going to make gene groups from the KEGG Pathways and the GO IDs.

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

3.1.2 In the next window, select **KEGG Pathways** from the *Name* pull down menu and click ***<OK>*** (see Figure 3-1).

Figure 3-1. Create a new grouping.

3.1.3 KEGG Pathways is selected as the text field in the next window (see Figure 3-2). The same row can contain multiple KEGG Pathways, separated by a "|" (e.g. 04060|04620|05120|04060|04620|05120). Use a "|" (a pipe) as delimiter. This will split the multiple IDs for a certain row entry.

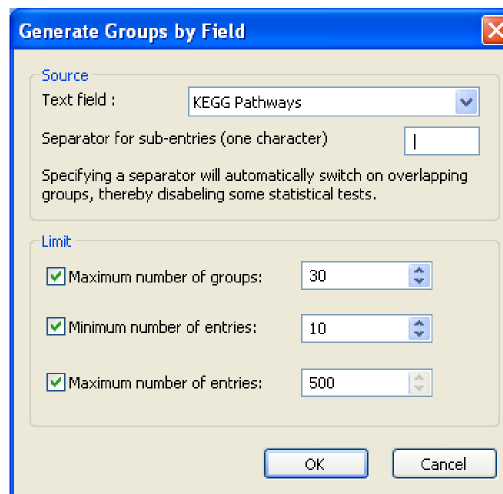


Figure 3-2. Creating groups based on the KEGG pathways.

3.1.4 Change the settings in the *Limit* panel as specified in Figure 3-2. This filtering is useful because we do not want small groups, this would make them useless in further analysis. Click ***<OK>***.

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

3.1.6 Press the ***<Create New Grouping>*** button once more and select **GO ID** from the *Name* pull down menu. Click ***<OK>***.

3.1.7 The same row entry can contain multiple GO IDs, separated by a "|" (e.g. GO:0006955|GO:0005529|GO:0007166). Use a "|" as delimiter. This will split the multiple IDs for a certain row entry.

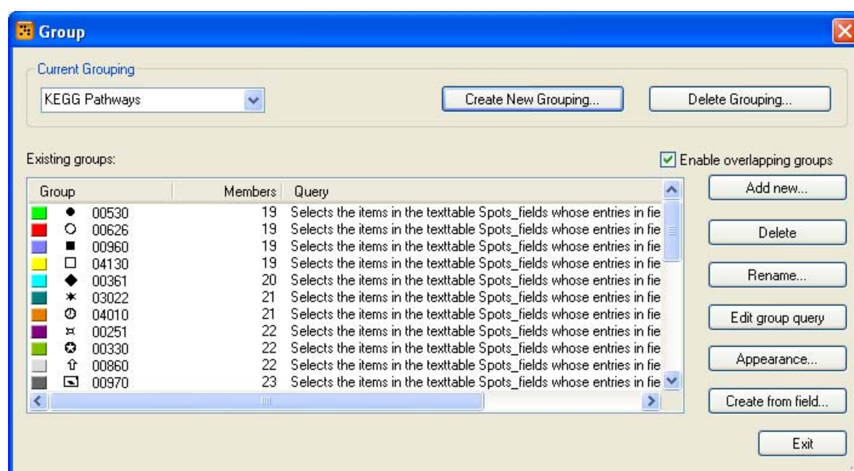


Figure 3-3. Groups based on the KEGG pathways.

3.1.8 Change the settings in the *Limit* panel as specified in Figure 3-4. Click **<*OK*>**.

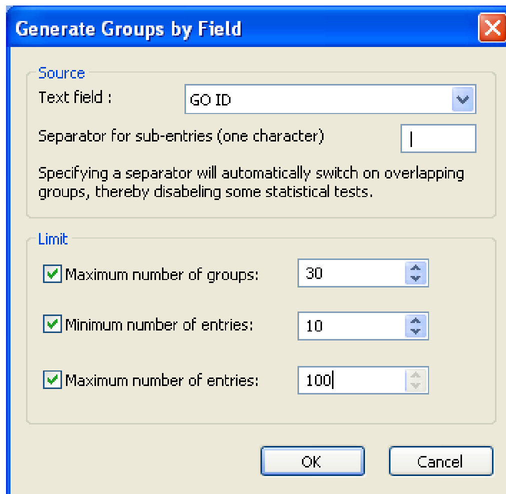


Figure 3-4. Creating groups based on the GO IDs.

3.1.9 The groups based on the settings are shown in the next window (Figure 3-5).

3.1.10 Press **<*Exit*>**.

GeneMaths XT offers the possibility to link a grouping to a website. We will link the GO ID grouping to the GO-website and the KEGG-pathways grouping to a KEGG-website. Later on we can use these links when using statistics reports.

3.1.11 Select ***Groups* > *Row Group Link***.

3.1.12 In the next dialog box, select **GO ID** and **http://www.godatabase.org/cgi-bin/amigo/go.cgi?action=replace_tree&query=###**. Click **<*OK*>**.

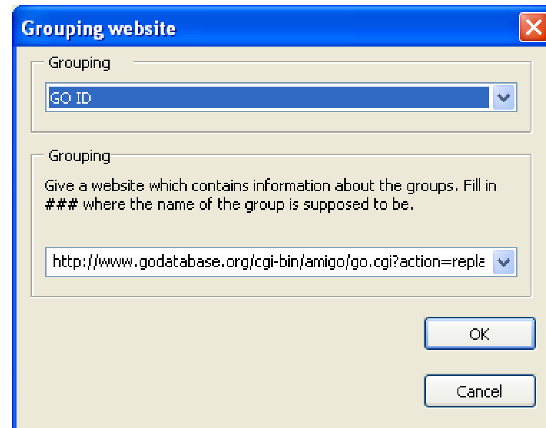


Figure 3-6. Linking GO ID grouping to a website.

3.1.13 Select ***Groups* > *Row Group Link*** once more. Select **KEGG Pathways** from the list and choose **http://www.genome.jp/dbget-bin/show_pathway?MAP ###** from the drop down menu. Click **<*OK*>**.

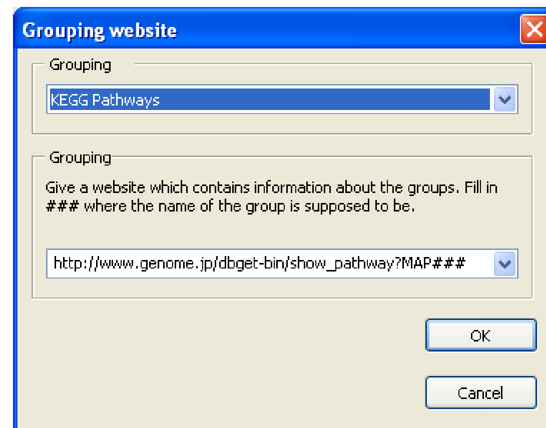


Figure 3-7. Linking 'KEGG Pathways' grouping to a website.

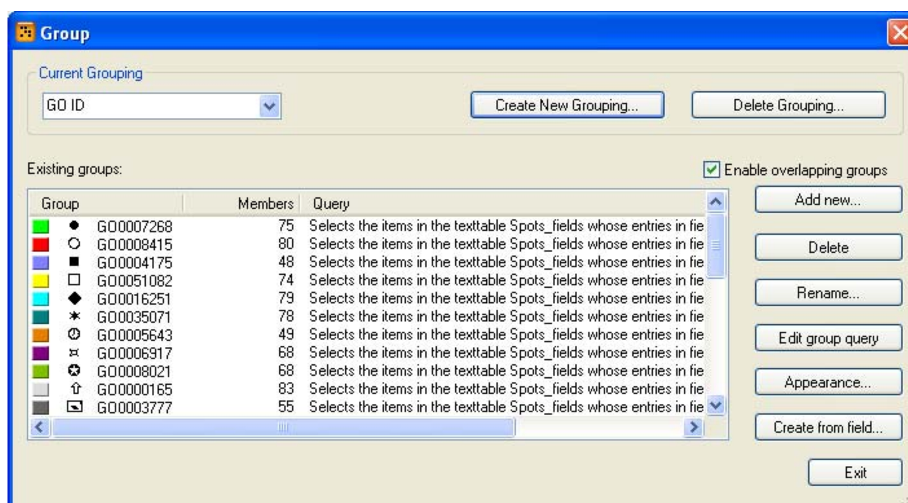


Figure 3-5. Groups based on the GO IDs.

3.2 Column groups

In the next step, we are going to make groups from the column information field **timepoint**.

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

3.2.2 Select **timepoint** from the pull down menu and press *<OK>*.

3.2.3 Uncheck all limits in the *Limit* panel. Press *<OK>*.

3.2.4 The groups based on the settings are shown in the next window (Figure 3-8).

3.2.5 Press *<Exit>*.

The groups and their colors are shown next to the column names:

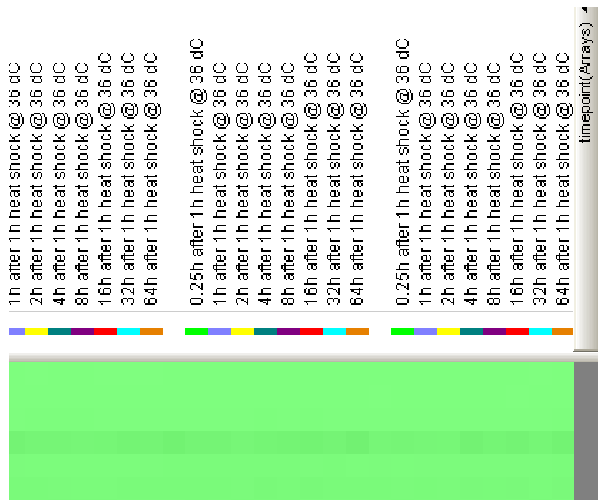


Figure 3-9. The timepoint groups.

3.2.6 Select the **timepoint** grouping from the pull down menu in the *Groups* window. All groups belonging to this grouping are listed.

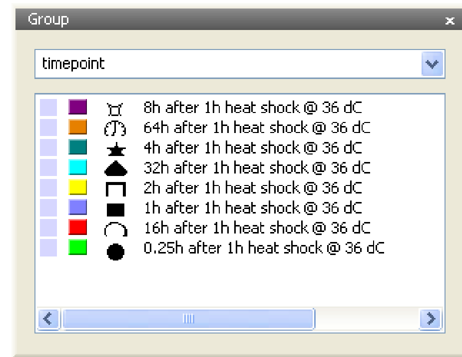


Figure 3-10. The *Group* window.

Next, we are going to assign all columns that are not assigned to a group in the **timepoint** grouping to a new group.

3.2.7 Select the entries in the *Column names* panel by CLTR clicking on them. Four entries are selected (see Figure 3-11).

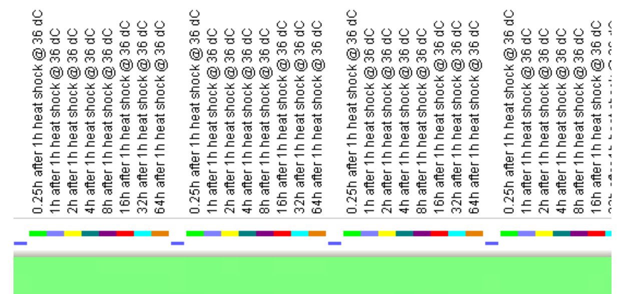


Figure 3-11. Column selection.

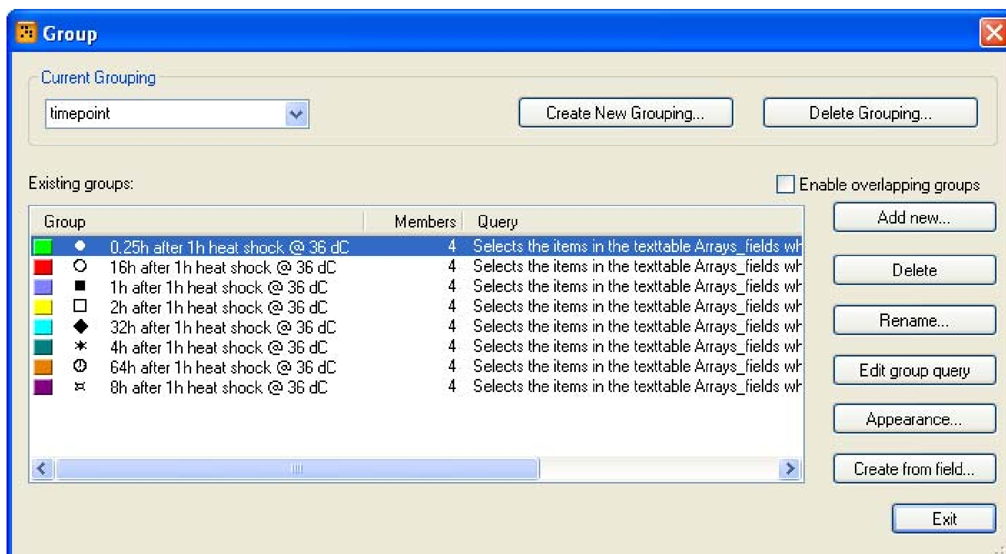


Figure 3-8. Grouping based on the time points.

3.2.8 Select ***Groups* > *Edit Column Group***.

3.2.9 Make sure the **timepoint** grouping is selected and press ***<Add new>***.

3.2.10 Name the new group **0h** and press ***<OK>***. Press ***<Yes>*** to confirm and press ***<Exit>***.

In a next step, we are going to group the columns based on their description: **control** versus **heat resistant.**

3.2.11 Make sure the **characteristics_ch1** is selected as column identifier. We are going to split this information content

3.2.12 Select ***Textfields* > *Split***. Fill out the settings as shown in Figure 3-12.

3.2.13 Press ***<OK>***.

The information in the characteristics_ch1 information field is split.

3.2.14 Select ***Groups* > *Edit Column Groups***.

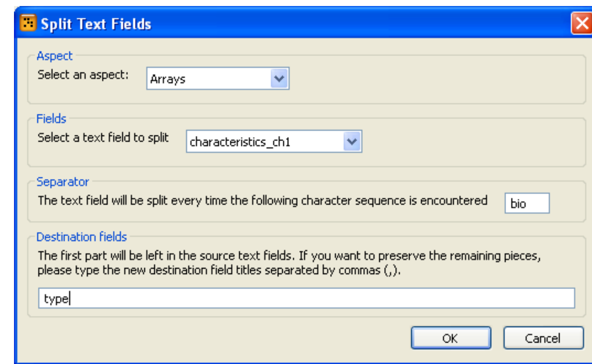


Figure 3-12. Split textfields.

3.2.15 Click on ***<Create New Grouping>***, select **characteristics_ch1** from the *Name* pull down menu and click ***<OK>***.

3.2.16 Uncheck all limits in the *Limit* panel and press ***<OK>***.

3.2.17 The groups based on the settings are shown in the next window (Figure 3-13).

3.2.18 Press ***<Exit>***.

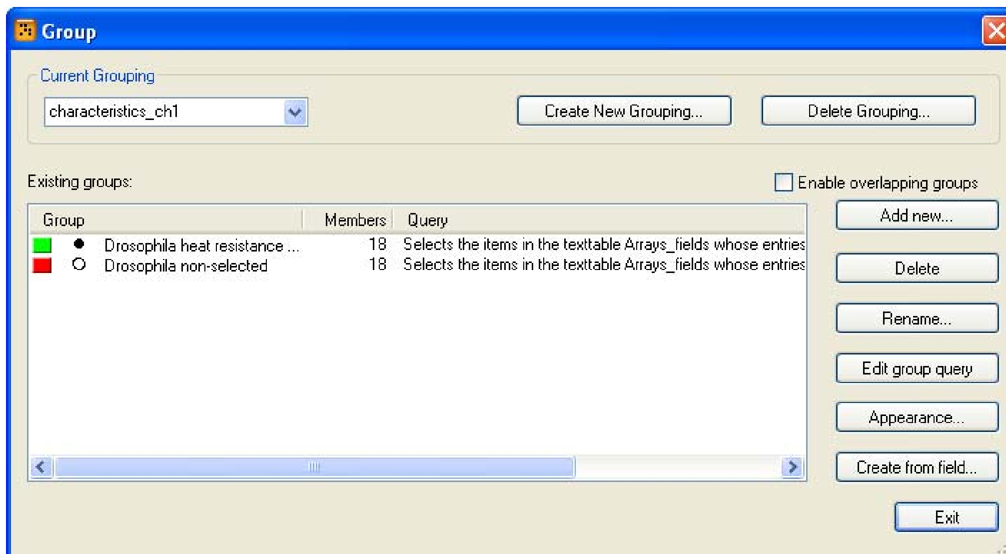


Figure 3-13. Grouping: control versus heat resistant.

4. Preprocessing

4.1 The preprocessing diagram

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

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

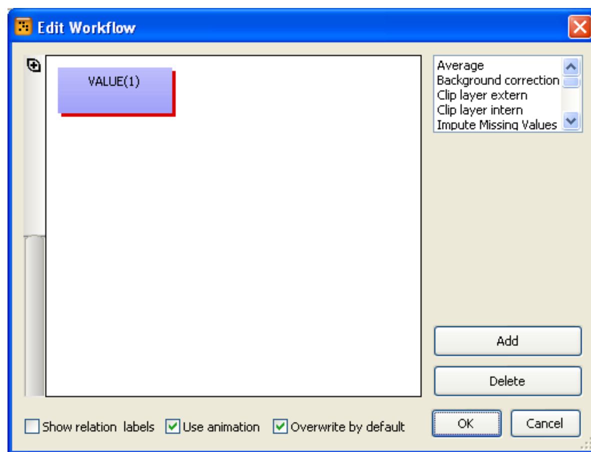


Figure 4-1. *Preprocessing* window.

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

4.1.2 Uncheck ***Overwrite by default***.

4.2 Log transformation

Before we can perform a statistical test we need to log transform our data.

4.2.1 Select the layer in the *Preprocessing* window.

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

4.2.3 Store the result of the log transformation in a new layer called **LogTrans**.

4.2.4 Press **<*OK*>.**

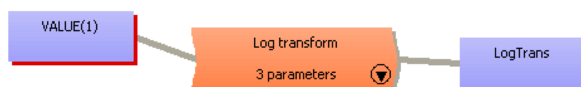


Figure 4-2. Log Transformation.

4.2.5 Click on the arrow in the Log Transform. Three parameters are displayed.

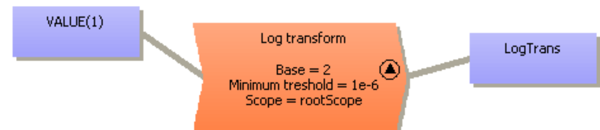


Figure 4-3. Parameters.

4.2.6 To change the parameters, double click in the orange box.

4.2.7 In this exercise we will use the default settings. Press **<*OK*>**.

4.2.8 Press **<*OK*>** to close the *Preprocessing* window.

The log2 intensities of the layer are calculated and stored. The new layer **LogTrans** is shown in the *Main* window (see Figure 4-4). Note that the histogram is not centered around zero yet.

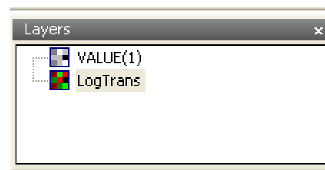


Figure 4-4. The LogTrans layer is added to the session.

4.3 Normalize arrays

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

4.3.1 In GeneMaths XT, select ***Layer > Preprocessing diagram*** again.

4.3.2 Make sure ***Overwrite by default*** is unchecked.

4.3.3 Select the **LogTrans** layer.

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

4.3.5 Store the result of the normalization in a new layer called **NormArrays**.

4.3.6 Click **<*OK*>**.

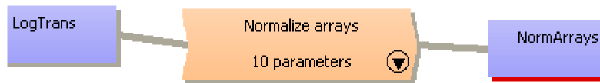


Figure 4-5. Normalize arrays

4.3.7 Press **<*OK*>** to launch the preprocessing tool.

The new layer is added to the list of layers in the *Layers* window. The histogram of the new layer NormArrays is now centered around zero as shown in Figure 4-6.

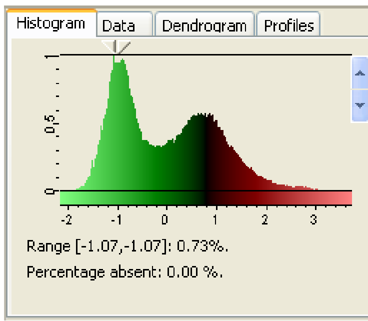


Figure 4-6. Data centered around zero.

4.4 Normalize genes

In order to perform some statistical analysis later on (e.g. creating a self-organizing map and a partitioning), we need to normalize our genes.

4.4.1 Select ***Layer* > *Normalization* > *Genes***. Fill out the dialog as shown in Figure 4-7.

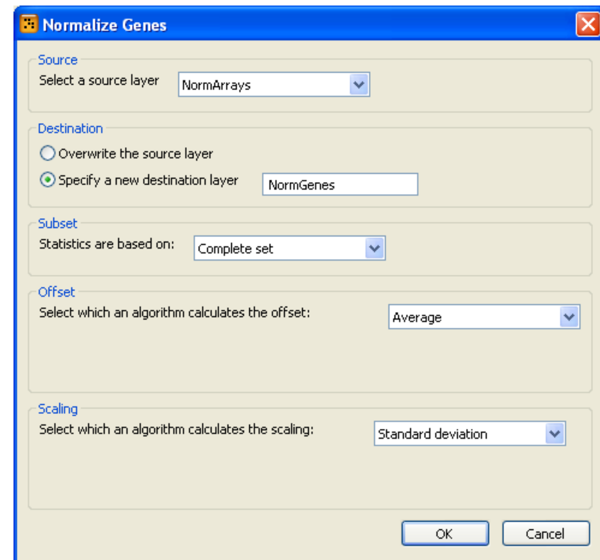


Figure 4-7. Settings for the normalization of the genes.

4.4.2 Press **<*OK*>**.

5. Statistics & Analysis

Now we have a starting point for further statistical analysis.

5.1 Statistical analysis on the 'timepoint' grouping

In this paragraph we are going to screen for differentially expressed genes between the 'timepoint' grouping. We will perform two statistical tests and look if these tests denote more or less the same genes as differentially expressed.

5.1.1 Select ***Profiles > Statistics Wizard***. Leave the orientation on **Row** and press **<*Next*>** (see Figure 5-1).

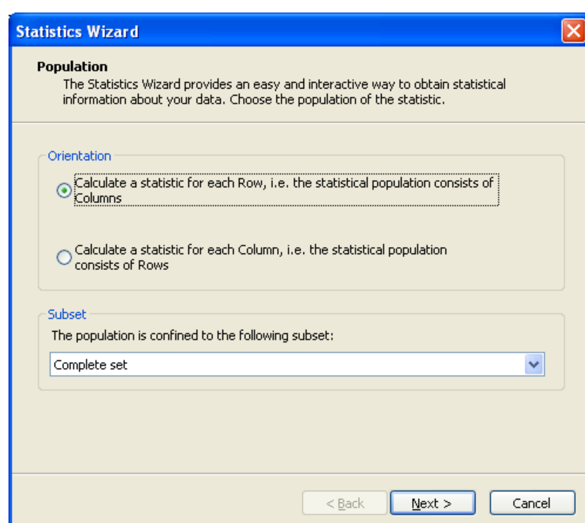


Figure 5-1. *Statistics* Wizard, step 1.

5.1.2 Select **ANOVA test** (under 'Independent test (multiple groups)') from the list and click **<*Next*>** (see Figure 5-2).

5.1.3 In the next window, make sure that **NormArrays** is selected and the **timepoint** grouping in the *Groups* panel. Select **p-value** as output and click **<*Next*>** (see Figure 5-3).

5.1.4 In the last window, choose the ***Benjamini & Hochberg procedure*** to correct for multiple testing and press **<*Finish*>**.

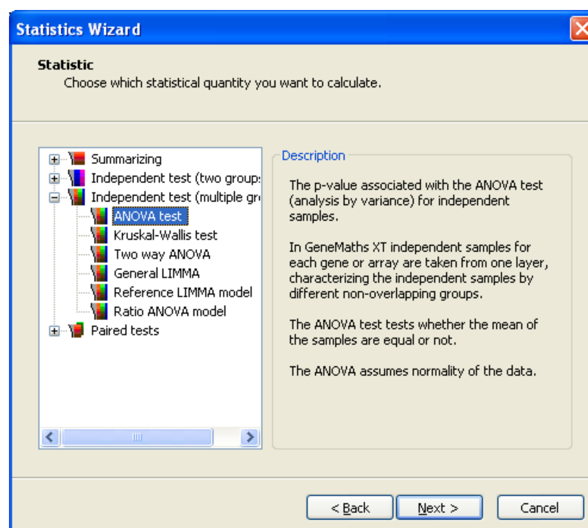


Figure 5-2. *Statistics* Wizard, step 2.

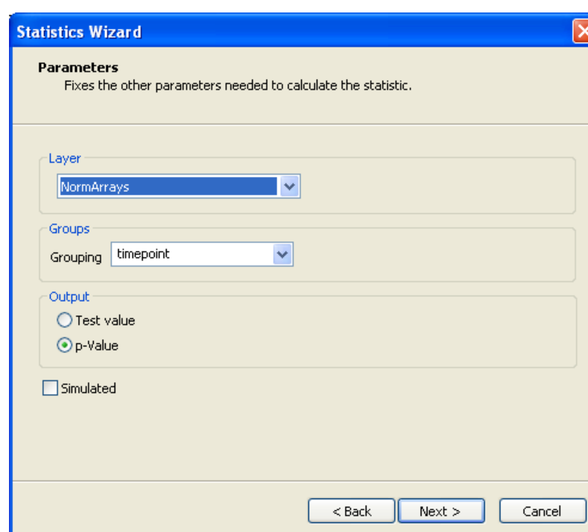


Figure 5-3. *Statistics* Wizard, step 3.

5.1.5 Click on the newly created profile in the *Profiles* tab. Right-click on the profile name and select ***Sort From Profile*** (see Figure 5-4).

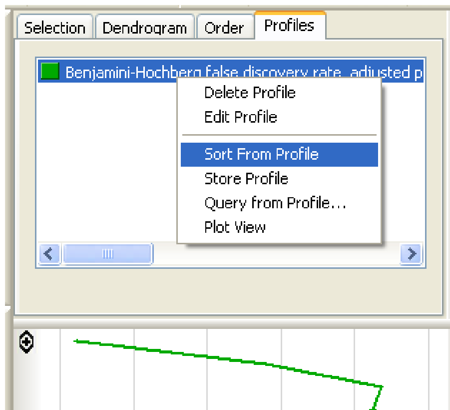


Figure 5-4. ***Sort From Profile.***

5.1.6 Right-click in the *Profile* panel (see Figure 5-5) and select ***Show as Numbers***.

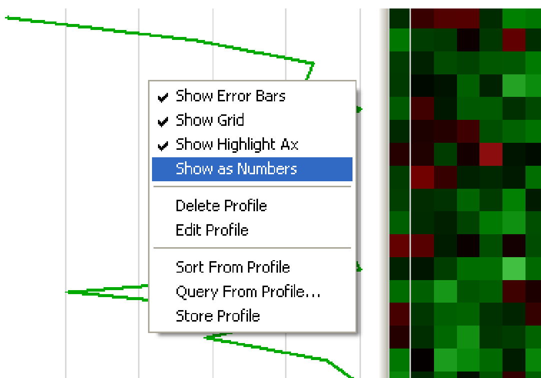


Figure 5-5. ***Show as Numbers.***

In the *Profile* panel, the p-values for the genes are shown. These p-values give an indication if genes are significantly differentially expressed in function of time or not. The lower the p-values, the more differentially expressed. In the next step we are going to select all genes with a p-value below 0.05.

5.1.7 Right-click in the *Profile* panel (see Figure 5-5) and select ***Query From Profile***.

5.1.8 Set the threshold of the p-values to '**< 0.05**' and press **<*OK*>**.

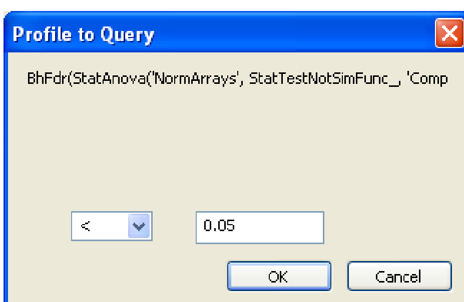


Figure 5-6. Setting a treshold.

The genes with a p-value smaller than 0.05 are selected in the *Main* window (blue arrow in the *Row names* panel).

5.1.9 Select ***Selection > Store Selection*** and store the selection as **ANOVA** (see Figure 5-7). Press **<*OK*>**.

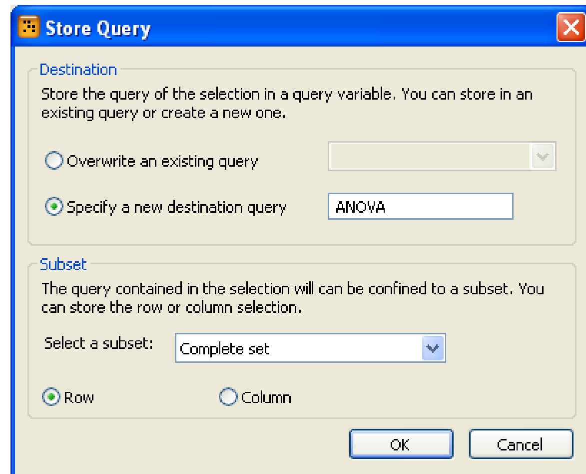


Figure 5-7. Storing the row selection.

We are going to perform an additional test and look if this test denotes more or less the same genes as differentially expressed.

5.1.10 First make sure that no entries are selected by pressing F4.

5.1.11 Select ***Profiles > Statistics Wizard***. Leave the orientation on **Row**. Press **<*Next*>**.

5.1.12 In the next window, select the **Reference LIMMA model** (under 'Independent test (multiple groups)') and click **<*Next*>** (see Figure 5-8).

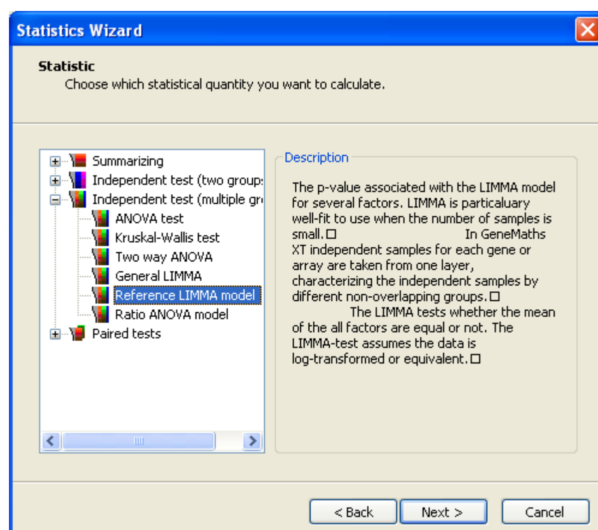


Figure 5-8. ***Statistics*** **Wizard, step 2.**

5.1.13 Fill out the settings as shown in Figure 5-3 and press **<*Next*>**.

5.1.14 In the next window, choose the ***Benjamini & Hochberg procedure*** to correct for multiple testing and press ***<Next>***.

5.1.15 Select the newly created profile in the *Profiles* tab and right-click on the profile name (see Figure 5-9). Select ***Query from Profile***.

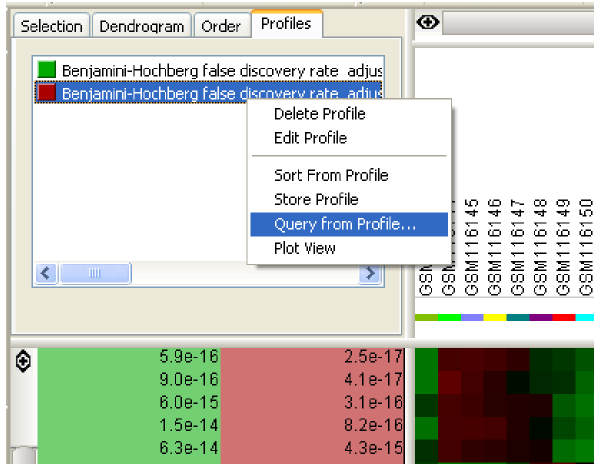


Figure 5-9. *Query from Profile*.

5.1.16 Set the threshold of the p-values to '**< 0.05**' and press ***<OK>***.

5.1.17 Select ***Edit > Store Selection*** and store the selection as **LIMMA**. Press ***<OK>***.

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

The Venn diagram is shown in a new window. 1125 row entries are shared amongst the two tests (see Figure 5-12).

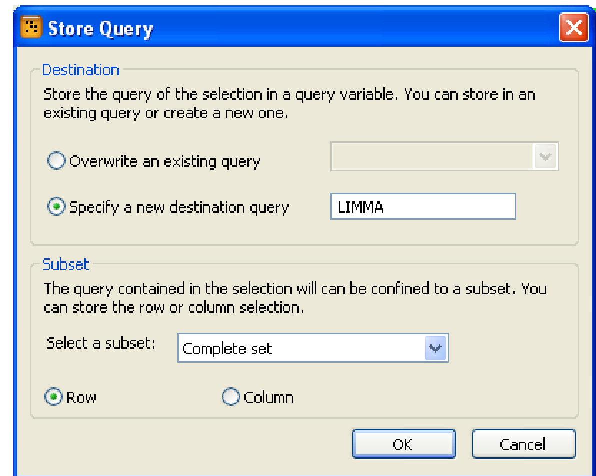


Figure 5-10. Store the row selection.

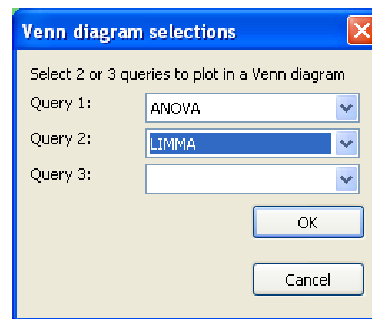


Figure 5-11. Select the queries to plot in a Venn diagram.

5.1.19 Click on the gene number '1125' in the Venn diagram. Select ***<Yes>*** to confirm that you want to select all the entries that are shared amongst the two tests.

5.1.20 Close the Venn diagram. The 1125 shared row entries are selected in the *Main* window.

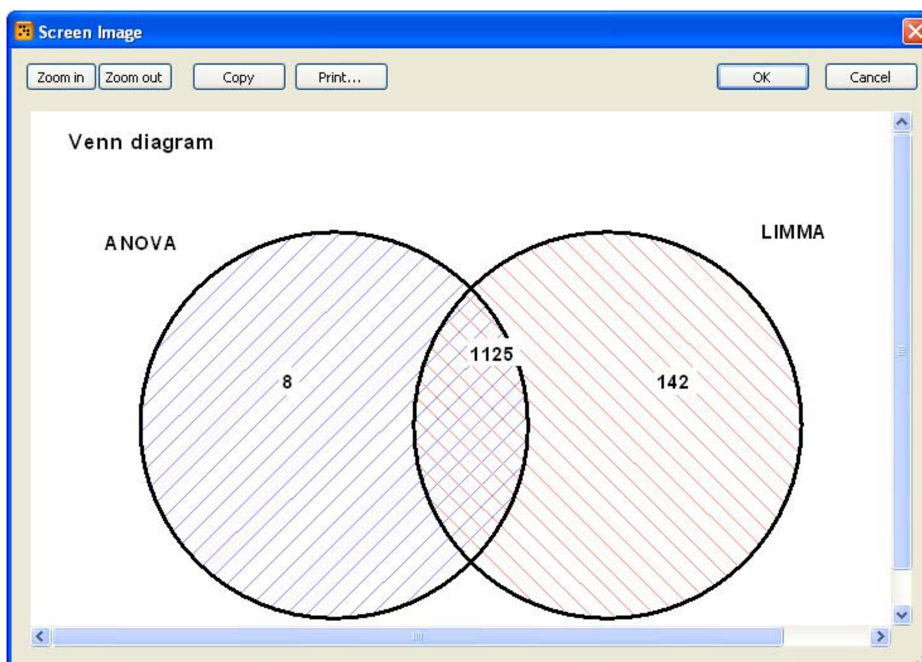


Figure 5-12. Venn diagram.

5.1.21 Select *Subset > Selection to Subset*, name the new subset **Diff expressed** and press *<OK>*.

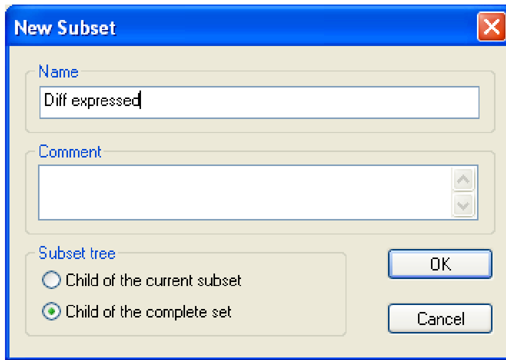


Figure 5-13. Creating a subset.

5.1.22 Press F4 to unselect all entries.

5.2 Statistical analysis: control versus heat resistant

In this paragraph we are going to screen for differentially expressed genes between the 'characteristic_ch1' grouping.

5.2.1 Make sure that no entries are selected by pressing F4.

5.2.2 Select the **Complete** subset and the **NormArrays** layer in the *Main* window.

5.2.3 Select *Profiles > Statistics Wizard*. Leave the orientation on **Row**, make sure that the **Complete** subset is selected and press *<Next>* (see Figure 5-14).

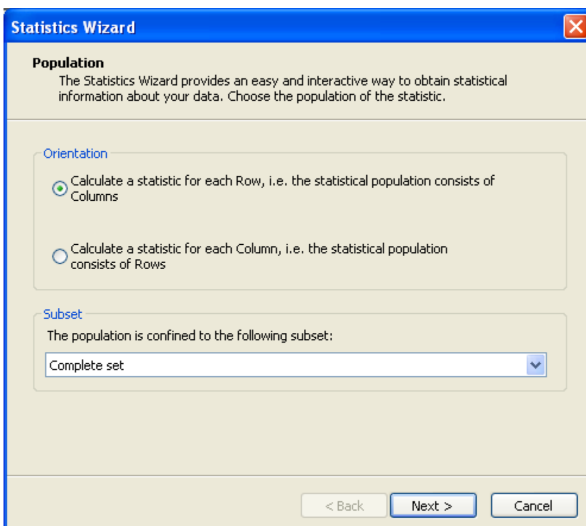


Figure 5-14. ***Statistics*** **Wizard, step 1.**

5.2.4 Select the **Independent t-test** (under 'Independent test (two groups)') and click *<Next>* (see Figure 5-15).

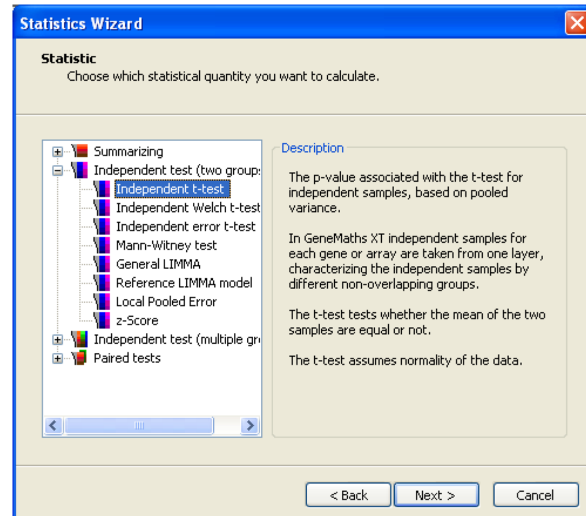


Figure 5-15. ***Statistics Wizard*****, step 2: selecting the test.**

5.2.5 In the next window, make sure that **NormArrays** is selected and the two different **other** groups. Select **p-value** as output and click *<Next>* (see Figure 5-16).

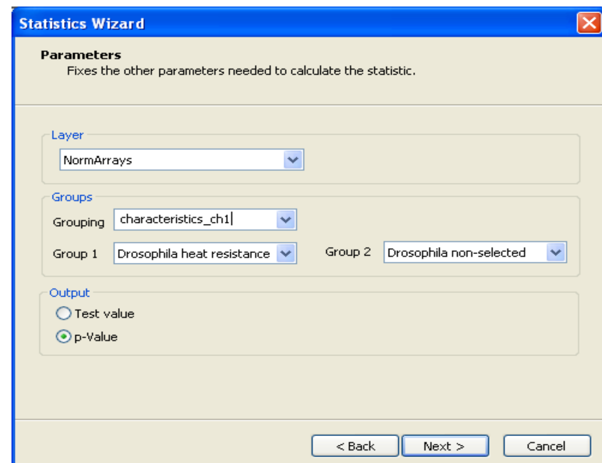


Figure 5-16. ***Statistics Wizard*****, step 3.**

5.2.6 In the final window, choose the ***Benjamini & Hochberg procedure*** to correct for multiple testing.

5.2.7 Click *<Finish>*.

5.2.8 Click on the newly created profile in the *Profiles* tab. Click right on the profile name and select *Sort From Profile* (see Figure 5-17).

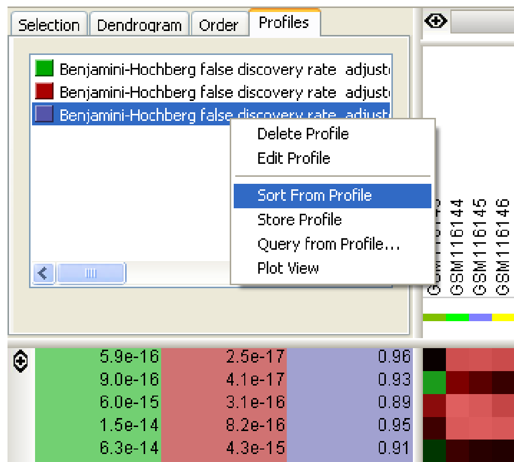


Figure 5-17. ***Sort From Profile.***

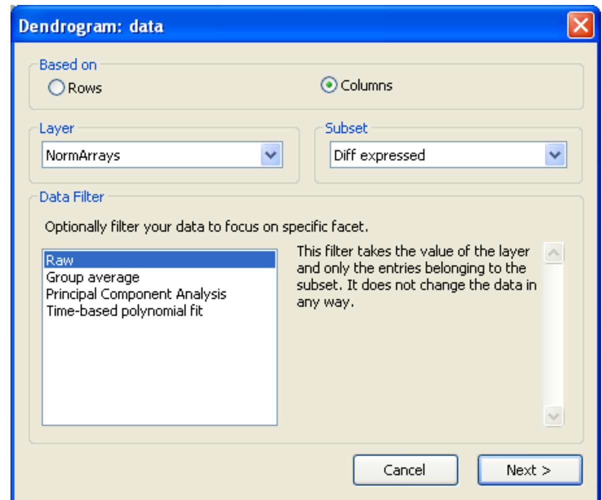


Figure 5-18. Cluster analysis: step 1.

5.2.9 Right-click on 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 the mean of the two groupings (control versus heat resistant) are the same. Most p-values have a value higher than 0.01, indicating that most genes are not differentially expressed between the groups.

5.3 Cluster analysis

In this paragraph we are going to cluster the columns based on the differentially expressed genes present in the subset **Diff expressed** (see Figure 5-13).

A) Hierarchical clustering

First we are going to create a UPGMA dendrogram, using a Euclidian distance similarity coefficient.

5.3.1 Select ***Analysis > Cluster Analysis***. Fill out the dialog box as shown in Figure 5-18 and press ***<Next>***.

5.3.2 Press ***<Next>*** once more and press ***<Finish>***.

The dendrogram based on the arrays is shown in the *Main* window of GeneMaths XT (see .

5.3.3 Select the **timepoint** column information field (see Figure 5-19).

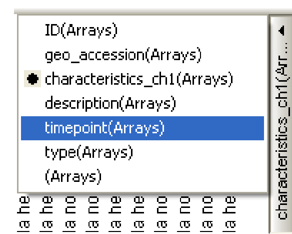
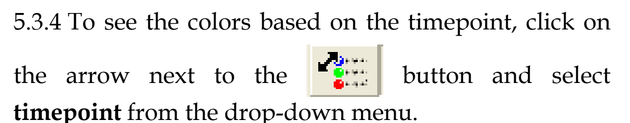


Figure 5-19. Selecting the timepoint information field

5.3.4 To see the colors based on the timepoint, click on the arrow next to the  button and select **timepoint** from the drop-down menu.

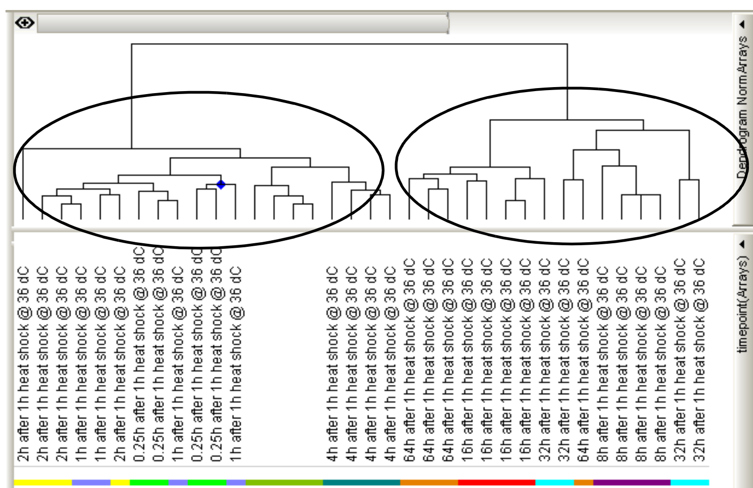


Figure 5-20. Dendrogram based on the arrays.

The dendrogram clearly shows a pattern with dependency on time. The late time points from 8 to 64 hours make up 1 major cluster, whereas the early time points (0 to 4 hours) make up another cluster (see Figure 5-20).

B) Dimensioning techniques

Next, we are going to perform a principal components analysis (PCA) on the same dataset to see if we get the same results.

5.3.5 Select ***Analysis > Principal Components Analysis*** or press the button.

5.3.6 Fill out the *PCA* dialog box as shown in Figure 5-21 and press ***<OK>***.

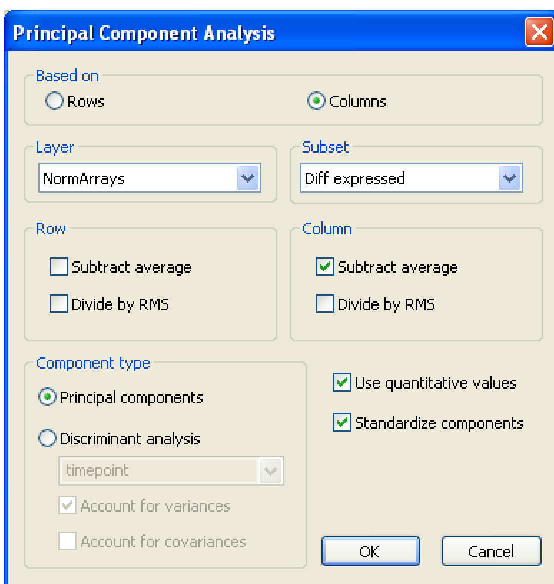


Figure 5-21. Dimensioning settings.

The PCA, based on the columns is shown in the PCA view (see Figure 5-22). The projection of the columns on the principal components is shown in the left panel with colors corresponding to the current grouping (in this case the **timepoint** grouping). The projection of the rows on the principal components is presented in the right panel.

5.3.7 Press F4 to unselect all entries.

5.3.8 Press the button in the toolbar.

5.3.9 Select all column (left panel) in the right part of the panel by dragging the mouse while holding the left button (see Figure 5-23).

The selected entries are surrounded by a blue square.

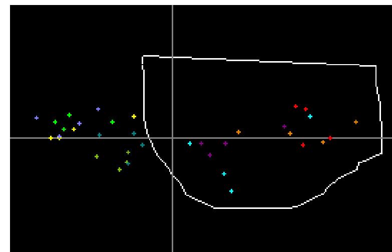
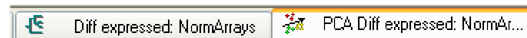


Figure 5-23. Selection of columns.

5.3.10 Return to the *Main* window by selecting the **'Diff expressed: NormArrays'** tab (see screenshot below).



The selected entries correspond to the members of one of the two clusters of our UPGMA dendrogram. From the results of the UPGMA clustering and the PCA we can conclude that there is a clear dependency on time.

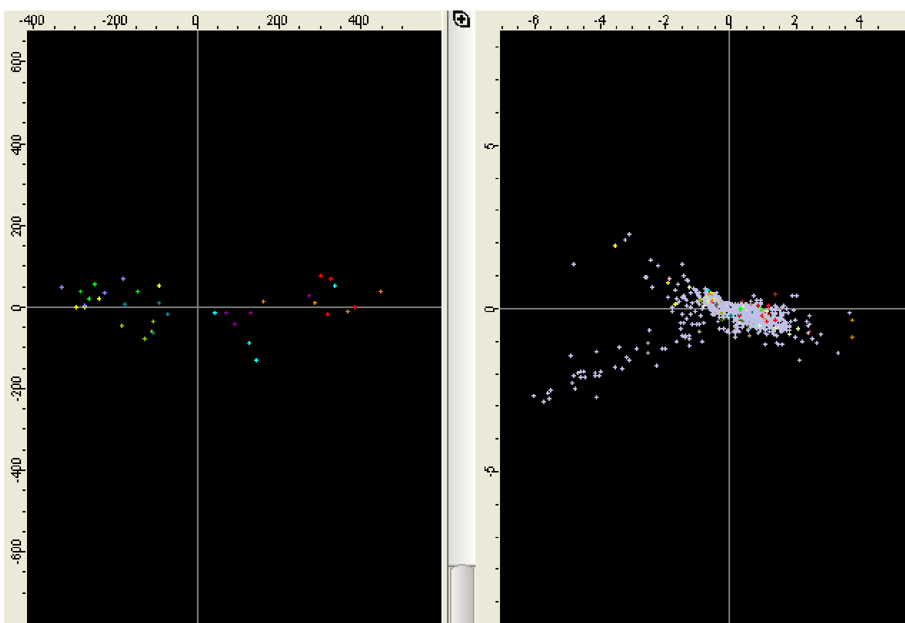


Figure 5-22. PCA view.

5.4 Creating a self organizing map (SOM) and a partitioning

In the next paragraph we are going to create a self organizing map and a partitioning. First we need to arrange our columns by increasing time points.

5.4.1 Select ***Textfields* > *To Profile***.

5.4.2 Select 'time' from the drop-down menu and press **<*OK*>** (see Figure 5-24).

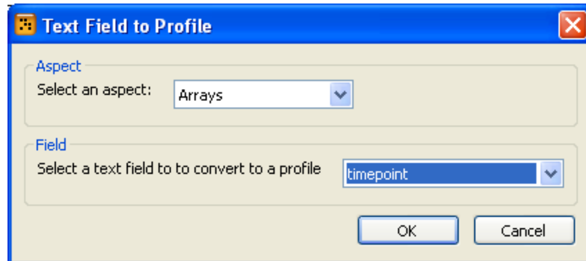


Figure 5-24. Convert to textfield.

5.4.3 Select the newly created profile in the *Profiles* tab. Right-click on the profile name and select ***Sort From Profile*** from the floating menu (see Figure 5-25).

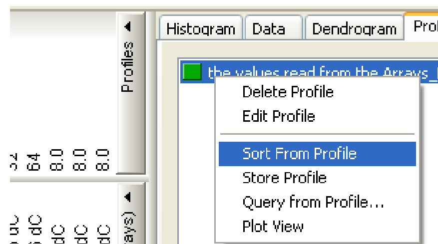


Figure 5-25. *Sort From Profile*.

5.4.4 The arrays are now arranged by increasing time points. Right click in the panel shown in Figure 5-26 and uncheck ***Show as numbers***.

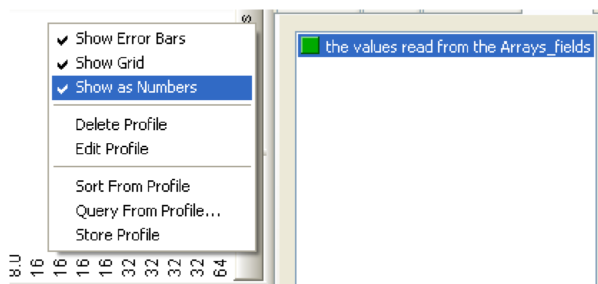


Figure 5-26. Profile shown as numbers.

A) Self-Organizing map

5.4.5 Select ***Analysis* > *Self-Organizing Map*** or press the button.

5.4.6 The *Self-Organizing Map* dialog box pops up. Fill out the settings as shown in Figure 5-27 and press **<*Next*>**.

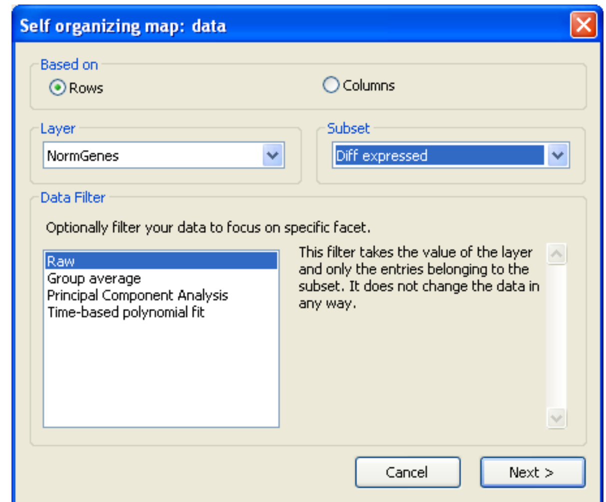


Figure 5-27. *Self-Organizing Map* dialog box.

5.4.7 Press **<*Next*>** once more.

5.4.8 Set the number of nodes in the X-dimension to **8**, and in the Y-dimension to **6.** Press **<*Finish*>**.

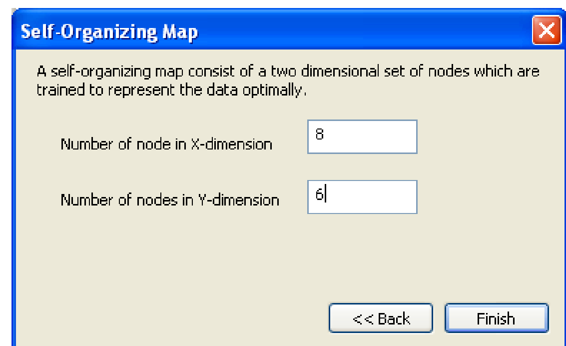


Figure 5-28. Dimensioning settings.

The result of the calculations is displayed in the SOM view. The expression values representation is the standard setting.

5.4.9 Select ***View* > *Cell Profile*** or press .

The cells are now represented by a profile that is the average of all gene profiles in the cell (see Figure 5-29). The error bars on the cell profiles are calculated from the standard deviation of the profiles in the cell from the average cell profile. We can clearly see a difference in cell profiles: the profiles on the left side show a downregulation of the genes, the profiles on the right indicate an upregulation (early and late).

5.4.10 Select *View* > ***Group Pie Chart*** or press .

The cells are represented by a circle where the groups of the current gene grouping (here: GO ID) are indicated as colored pie charts (see Figure 5-30). Genes with no associated GO IDs are represented in gray.

5.4.11 Select a cell (preferably one with a big colored piece of pie, for example the one surrounded by the yellow square in Figure 5-30), right-click in the cell and select ***Cell to Statistics Report*** from the floating menu.

The report (see Figure 5-31) gives an indication if the number of GO IDs in the selected cell is more or less than predicted by chance.

From this statistics report in Figure 5-31 we can see that GO:0051082 has 1.5% (0.015) chance of being present in the cell. The actual ratio of being present in the selected cell is 40% (= 0.40 = 4/10). This means that the odds of having that many number of GO:0051082 in the selected cell is much more than by chance.

5.4.12 Click on the GO number in the statistics report.

5.4.13 The GO-website for the selected GO numberopens (see Figure 5-32).

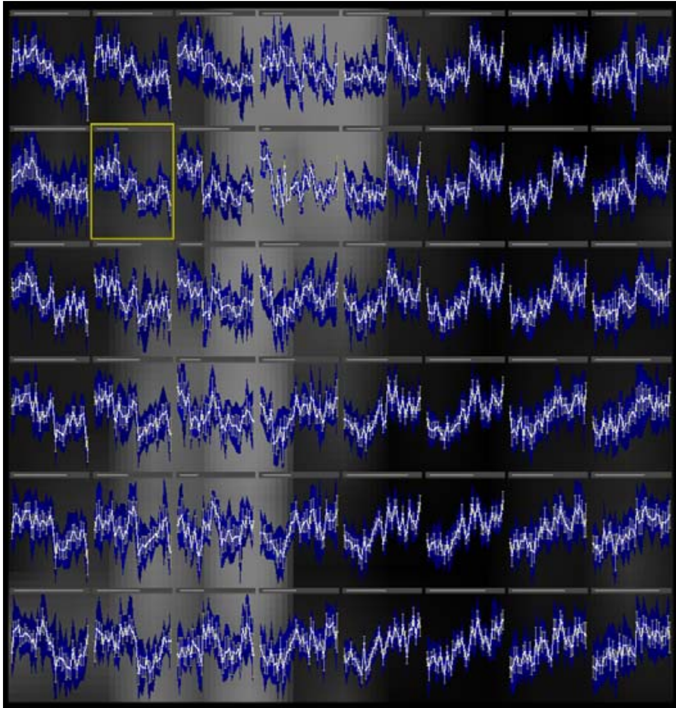


Figure 5-29. SOM view: Cell profile.

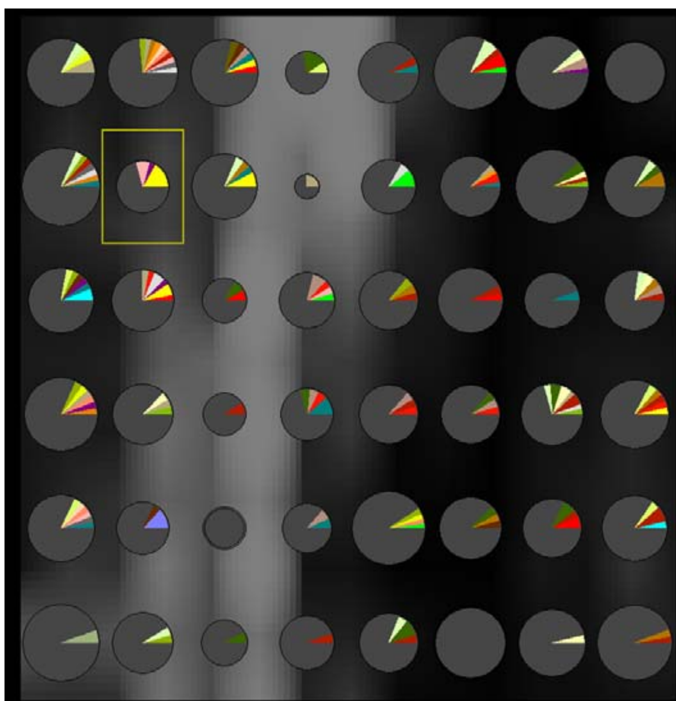


Figure 5-30. SOM view: Group pie chart (GO ID grouping).

Statistics Information

Statistics report

Grouping :GO ID
Subset: Diff expressed

Total number of items: 1125
Selected number of items: 17
Expected ratio: 0.0151111

Category	*Binomial p-Value*	*Hypergeometric p-Value*	*Total*	*Partial*	*Ratio*	*z-Score*
GO00510820	7.1188e-006	10	4	0.4	10.017	
GO00426231.58007e-012	0.00209412	5	2	0.4	7.06729	
GO00069170.000686514	0.0734331	5	1	0.2	3.39491	
GO00059760.586236	0.747022	19	0	0	-0.544298	
GO00450870.617916	0.782489	16	0	0	-0.498806	
GO00350710.679856	0.845146	11	0	0	-0.412659	

Copy to clipboard OK

Figure 5-31. Statistics report.

5.4.14 Close the statistics report by clicking **<*OK*>** or select **<*Copy to clipboard*>**.

5.4.15 Select ***Groups* > *Edit Row Groups*** and select ***KEGG Pathways*** from the drop-down menu (left upper corner). Press **<*Exit*>**.

5.4.16 Select one of the cells. Right-click in the cell and select ***Cell to Statistics Report***. Since we also linked the KEGG pathways to a website (see chapter 'Groupings'), you can click on one of the identifiers from the list. The corresponding pathway will open in your browser.

5.4.17 Close the statistics report.

B) Partitioning

Next, we are going to perform a partitioning and see if we get the same trends as with the SOM (downregulation versus upregulation).

5.4.18 Select ***Analysis* > *Partitioning***. The *Partitioning* dialog box pops up.

5.4.19 Fill out the settings as shown in Figure 5-27 and press **<*OK*>**.

5.4.20 Leave the settings in the next step unaltered and press **<*Next*>**.

The complete data set is shown as one cell. The expression values representation is the standard setting. You can split the cell manually or automatically. In this example we are going the split the complete data set manually.

5.4.21 Select the cell (now surrounded with a red square) and select ***Partitioning* > *Split cell***.

5.4.22 Set the number of partitions to 8 and press **<*OK*>**.

AmiGO: Gene Ontology Browser - Windows Internet Explorer

the Gene Ontology AmiGO

Advanced Search BLAST search Browse Help

Search GO Terms Genes or proteins Exact Match Submit Query

Filter tree view

all : all [238365]
GO:0008150 : biological_process [158738]
GO:0005575 : cellular_component [160999]
GO:0003674 : molecular_function [161397]
GO:0005488 : binding [45799]
GO:0005515 : protein binding [21864]
GO:0051082 : unfolded protein binding [411]

Figure 5-32. GO-website.

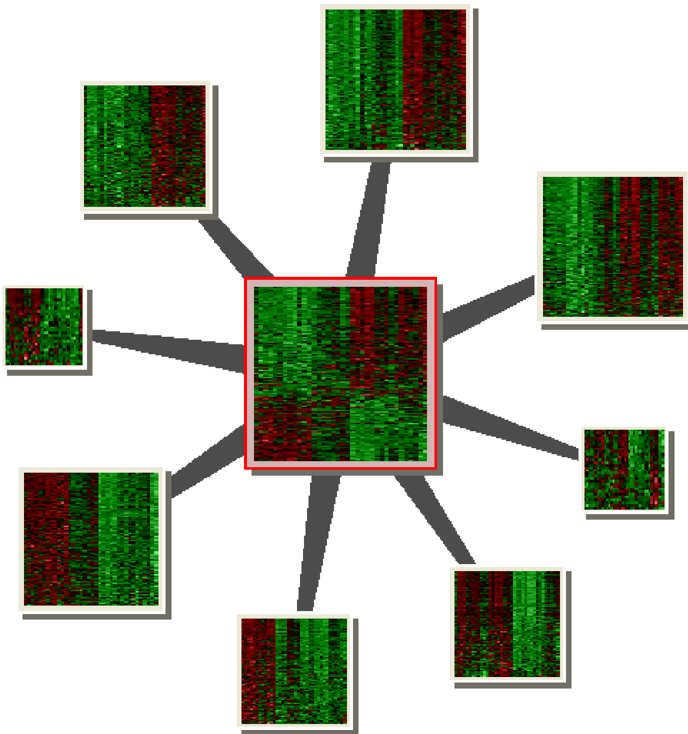


Figure 5-33. The *Partitioning* view.

The result of the calculations is displayed in the Partitioning view (see Figure 5-33).

5.4.23 Select ***View > Cell Profile*** or press .

The cells are now represented by a profile that is the average of all gene profiles in the cell. The error bars on the cell profiles are calculated from the standard deviation of the profiles in the cell from the average cell profile.

Just like with the SOM analysis, we can clearly see a difference in cell profiles: early-upregulated, late-upregulated and early-downregulated genes. In conclusion: from the results of SOM and Partitioning we can conclude that there are different categories in the 'Diff expressed' subset.