

GeneMaths XT

One color 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	3.3 Group window	16
1.1 Downloading the data	5		
1.2 Importing the data in GeneMaths XT	5	4. Preprocessing	19
2. Annotation	11	4.1 Log transformation	19
2.1 Spot information	11	4.2 Analyze groups	19
2.2 Spot annotation	11	4.3 Filtering	21
2.3 Array annotation	12	4.4 Normalization of arrays	23
3. Groupings	15	5. Statistics & Analysis	25
3.1 Row groups	15	5.1 Hierarchical clustering	25
3.2 Column groups	16	5.2 Differentially expressed genes	28
		5.3 Comparing statistical tests	32

1. Import

1.1 Downloading the data

An example dataset will be used in order to explain the workflow of GeneMaths XT. This dataset is publicly available on the GEO website ('Gene Expression Omnibus').

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 **GDS724**.

1.1.2 Press <Go>.

1.1.3 Select **GSE1563** 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 **GSE1563_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.

1.2 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>.

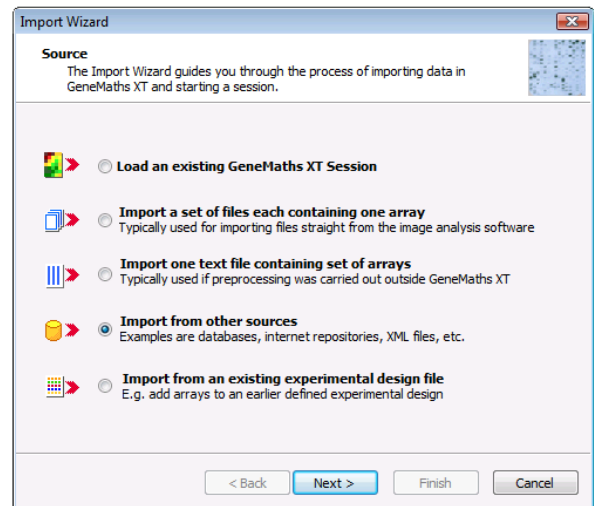


Figure 1-2. Import wizard: select data source.

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).

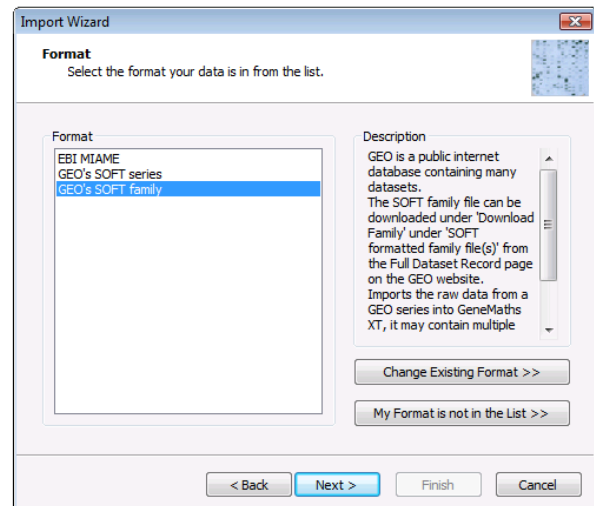


Figure 1-3. Import wizard: select format.

1.2.5 Click <Next>.

1.2.6 Browse for the stored file in the *File* panel. You can leave the top two panels empty (see Figure 1-4).

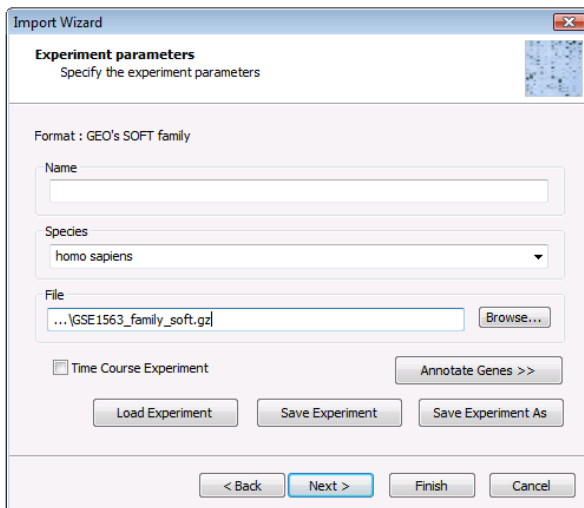


Figure 1-4. Start Wizard: input file.

1.2.7 Click <Next>.

1.2.8 Specify the name of the processed file e.g. GDS724.xps.

1.2.9 The *Calculation* dialog box pops up. The status of the import of the data is shown (see bottom of the box).

1.2.10 After the processing of the data (this may take a couple of minutes) GeneMaths XT will prompt to specify the contents of the columns in the *Define Format* dialog box (see Figure 1-5).

1.2.11 Select the second column **ID_REF** by clicking on it. The column is highlighted in pink. Specify the kind of data by changing the settings in the *Column information* panel. Select **Text** in the *Type* box and **Gene ID** in the *Text* box (see Figure 1-6).

1.2.12 Select the third column **VALUE**. Select **Quantitation** in the *Type* box and **Target, Foreground** and **Value** in the *Quantitation* box (see Figure 1-7).

1.2.13 Select the fourth column. Select **Quantitation** in the *Type* box and **Target, Foreground** and **Quality control** in the *Quantitation* box (see Figure 1-8).

1.2.14 Press the <Next & Copy> button. The fifth column now contains the same settings as the fourth column.

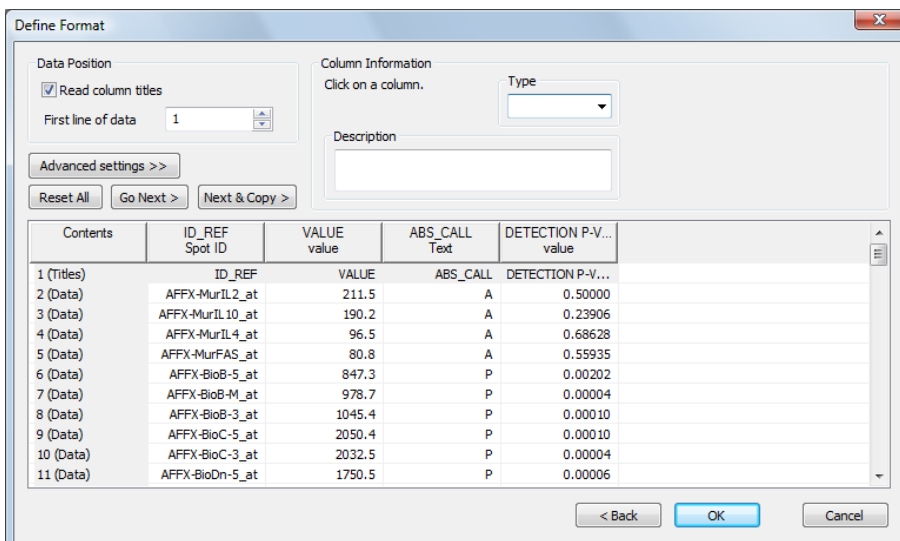


Figure 1-5. The *Define Format* dialog box.

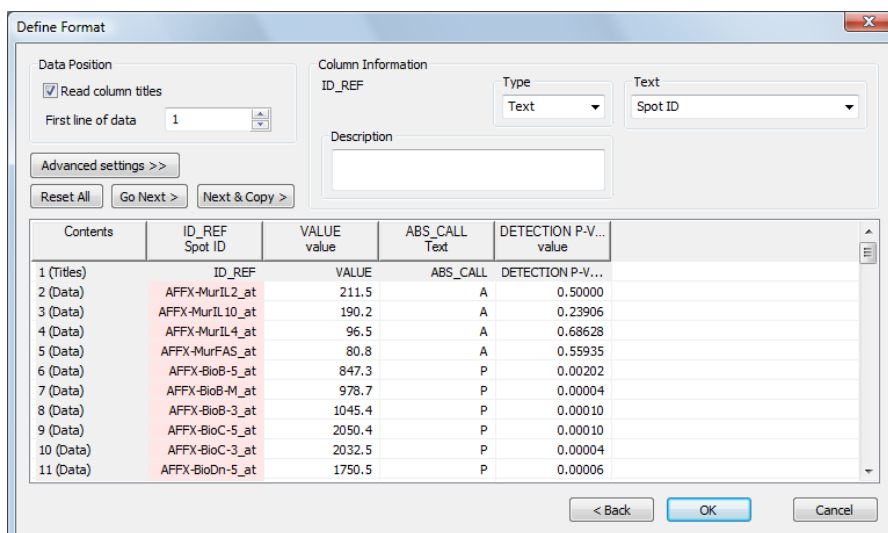


Figure 1-6. Settings for the second column.

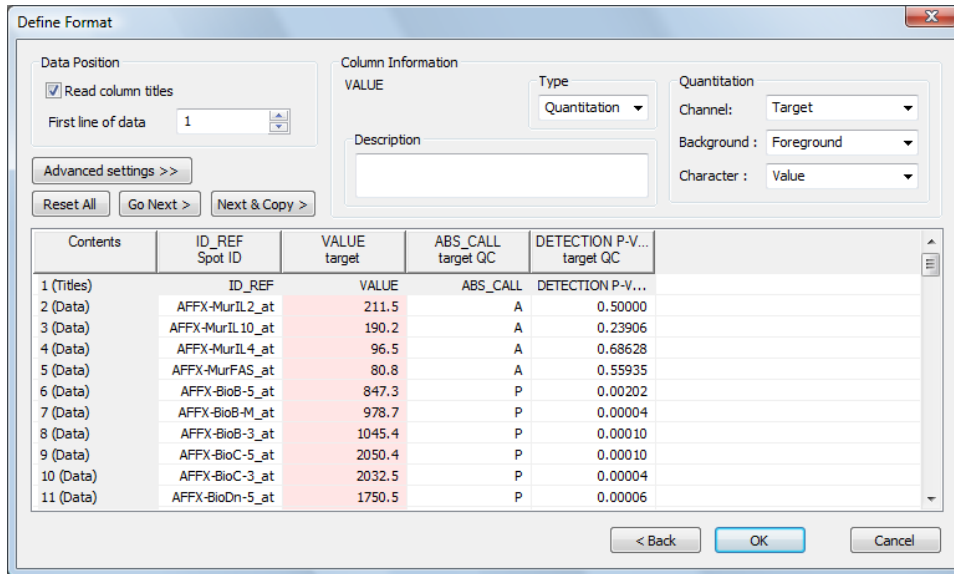


Figure 1-7. Settings for the third column.

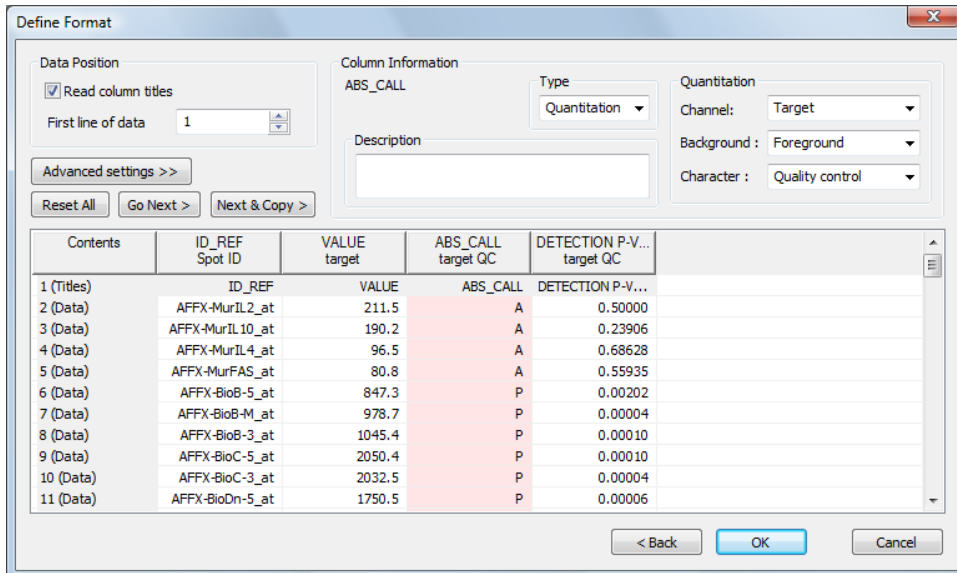


Figure 1-8. Settings for the fourth column.

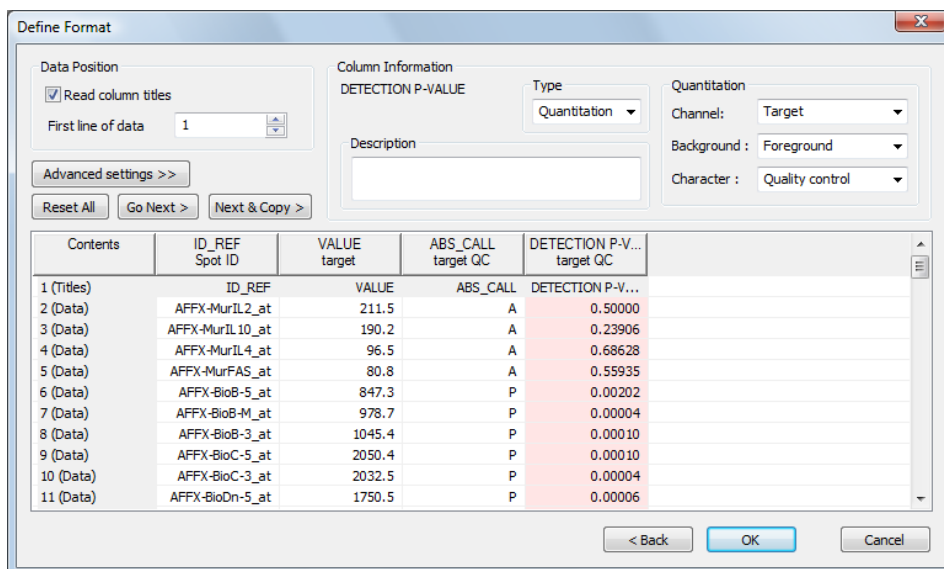


Figure 1-9. Settings for the last column.

1.2.15 After having specified the correct column information for all 4 data columns press **<OK>**.

1.2.16 The *Import mapping* dialog box pops up asking you to create a mapping for your data. This mapping tells GeneMaths XT which quantitations to use in the session.

1.2.17 Select **ID-REF** and hit “>”. ID-REF is now placed in the right box (see Figure 1-10).

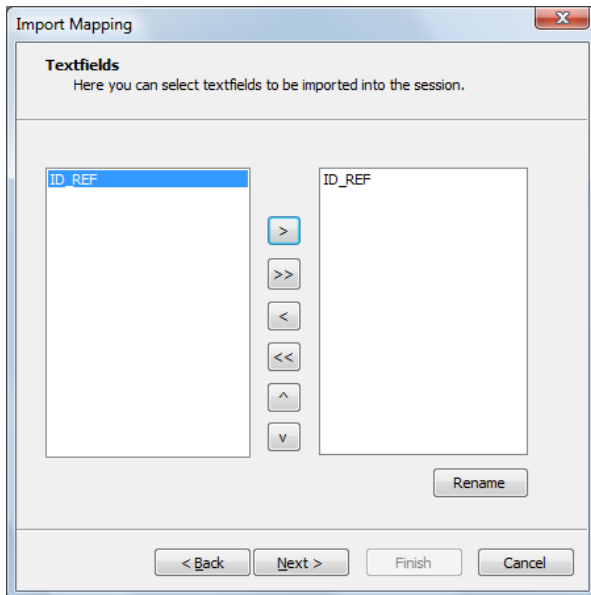


Figure 1-10. *Import mapping* dialog box, step 1.

1.2.18 Click **<Next>**.

1.2.19 Uncheck the two checkboxes in the *Filter* panel of the next dialog box. In the *Quantitations* panel, select ‘VALUE’ in the *Target* box (see Figure 1-11) and press **<Next>**. Entering these settings only loads the ‘VALUE’-column as a target signal.

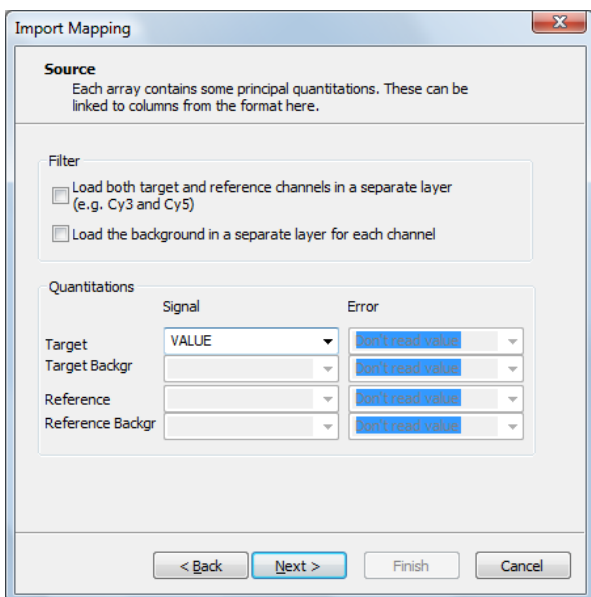


Figure 1-11. *Import mapping*, step 2.

1.2.20 In the next step of the import, select **DETECTION P-VALUE** in the first column, select **No error** in the second column and hit “>” (see Figure 1-12). ‘DETECTION P-VALUE’ is now used as an extra quantitation.

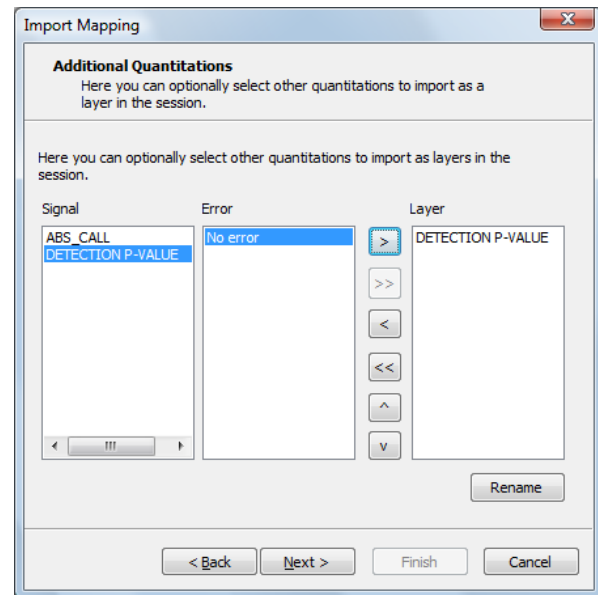



Figure 1-12. *Import mapping*, step 3.

1.2.21 Click **<Next>**.

1.2.22 In the final step of the *Import* wizard you can select which quality control criteria you want to use as a present/absent indication. In this example we could use the Affymetrix PMA-calls, but because we want to keep our value layer intact and have more flexibility during the analysis we will leave the settings unchanged.

1.2.23 Select **<Finish>**.

GeneMaths XT will import the data in a new session. The *Main* window of GeneMaths XT appears as depicted in Figure 1-13. The session contains two layers: ‘Target’ (imported from column ‘VALUE’) and ‘DETECTION P-VALUE’ (imported from column ‘DETECTION P-VALUE’). The layers are displayed in the top left panel.

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

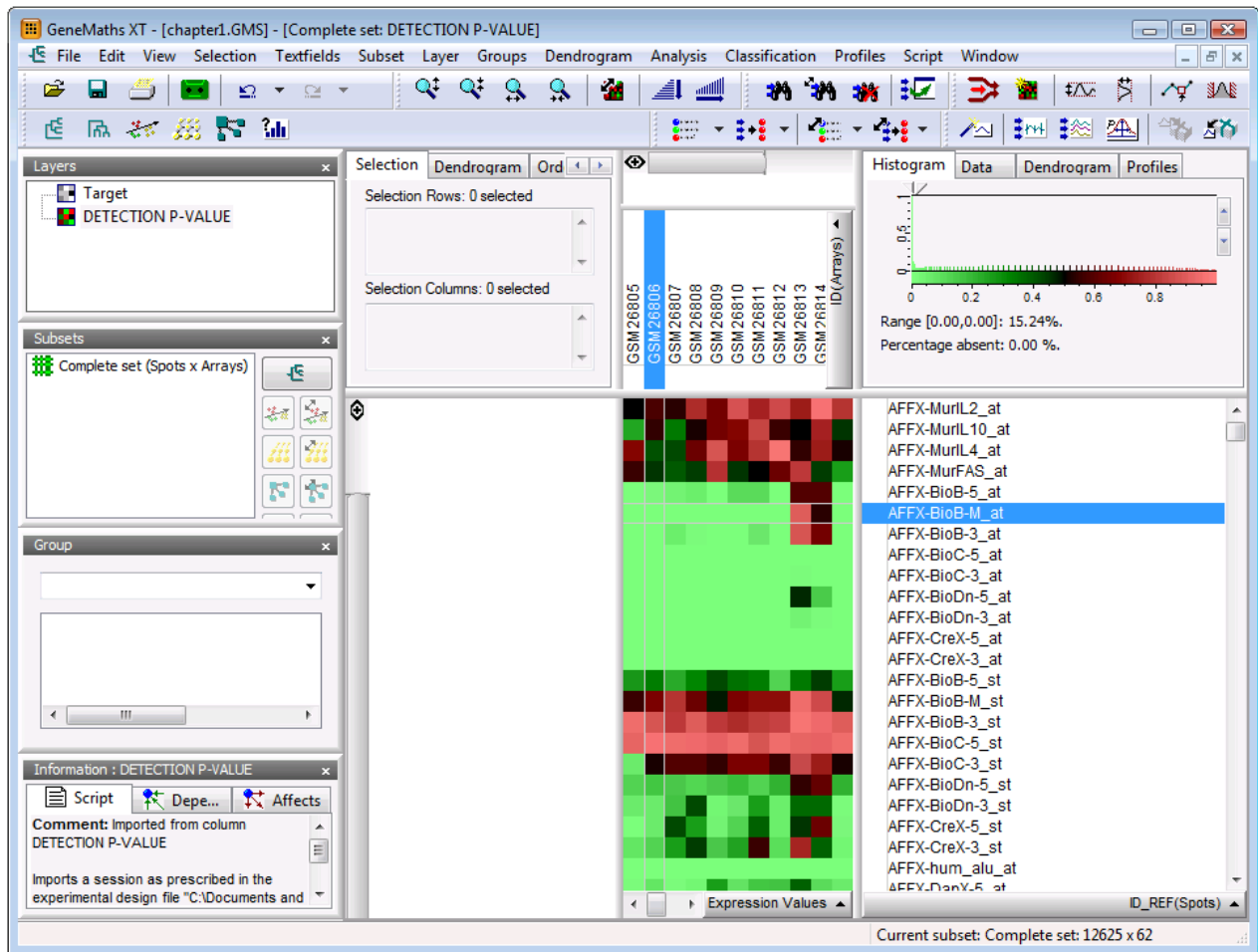


Figure 1-13. The *Main* window of GeneMaths XT after import of the data.

2. Annotation

2.1 Spot information

The annotation is automatically done with the import of the file. The spots however, do not yet contain the GO ID as an identifier. Because this experiment is done on an Affymetrix platform, we can directly use an annotation file based on this specific platform.

The first thing to do is to find out what kind of array GPL91 is and then use the corresponding *.gaf file.

2.1.1 Go to the GEO homepage: <http://www.ncbi.nlm.nih.gov/geo/>, click in the box next to 'Query > GEO accession' and type 'GDS724'.

2.1.2 Press <Go>.

Detailed information on the dataset is shown in the *GDS Summary* panel (Figure 2-1).

2.1.3 Click on the information in the *Platform* box. The next page on the website gives full detail about the used platform (see Figure 2-2). For this dataset we need the file 'Affymetrix_HG_U95Av2.gaf'.

2.2 Spot annotation

2.2.1 Select *Textfields > Annotations > Affymetrix* in the *Main window* of GeneMaths XT and select HG_U95Av2 from the list. Leave 'GeneMaths GAF file' enabled and make sure that GeneMaths XT will check for updates on the website (see Figure 2-3).

2.2.2 Press <OK>.

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

2.2.3 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 and press <OK>.

2.2.4 In the next window, use **Probe set ID** as a link for **ID_REF** and select **GO ID** as information field to add to the session.

GDS Summary			
Accession:	GDS724 View Expression (GEO profiles)		
Title:	Kidney transplant rejection expression profiling		
DataSet type:	gene expression array-based (RNA / in situ oligonucleotide)		
Summary:	Analysis of kidney transplant rejection by expression profiling of kidney biopsies and peripheral blood lymphocytes from patients. Results identify expression profiles unique to rejection, dysfunction without rejection, and well-functioning transplants.		
Platform:	GPL91: Affymetrix GeneChip Human Genome U95 Set HG-U95A		
Sample organism:	Homo sapiens	Platform organism:	Homo sapiens
Feature count:	12651	Value type:	count
Series:	GSE1563	PubMed ID:	15307835
Series published:	07/14/2004	Last GDS update:	09/21/2004

Figure 2-1. Array information on the GEO website.

Platform GPL91		Query datasets for GPL91
Status	Public on Mar 11, 2002	
Title	Affymetrix GeneChip Human Genome U95 Version [1 or 2] Set HG-U95A	
Technology type	in situ oligonucleotide	
Distribution	commercial	
Organism(s)	Homo sapiens	
Manufacturer	Affymetrix	
Manufacture protocol	see manufacturer's web site	

Figure 2-2. Detailed array information.

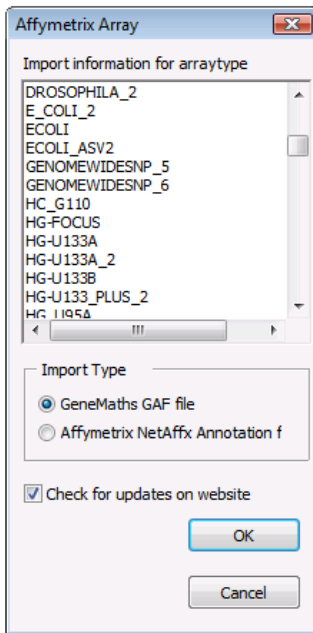


Figure 2-3. Select Affymetrix Arraytype.

2.2.5 Click on the bottom right tab in the *Main* window of GeneMaths XT (see Figure 2-4). GO ID is added to the list of row identifiers.

2.2.6 Select GO ID from the list. The available GO IDs are shown next to the spots (see Figure 2-4).

2.3 Array annotation

Next we want to add extra information to the arrays of our session. A file called 'experimental.txt' contains additional array information for this dataset and can be found on our website.

2.3.1 Select 'Download' in the menu bar of our home page (www.applied-maths.com) and select 'Manuals & tutorials'. Select 'GeneMaths XT Tutorials'. In the new window, click with your right mouse button on 'Array Annotations' next to 'One color design tutorial' and select 'Save Target As...'. Save the file on your computer.

2.3.2 Select *Textfields > Import*.

2.3.3 In the *Import Text Fields* dialog box select the aspect **Arrays** and navigate to the the experimental.txt file.

2.3.4 Use the **ID** information in the session and the **ID** column in the text file to link the information. Select all information fields except for ID (see Figure 2-5). To select all fields, select the first field, hold the SHIFT button and select the last field.

2.3.5 Press <OK>.

2.3.6 Click on the column identifiers tab (see Figure 2-6). The information fields are added to the list of column identifiers (red rectangle).

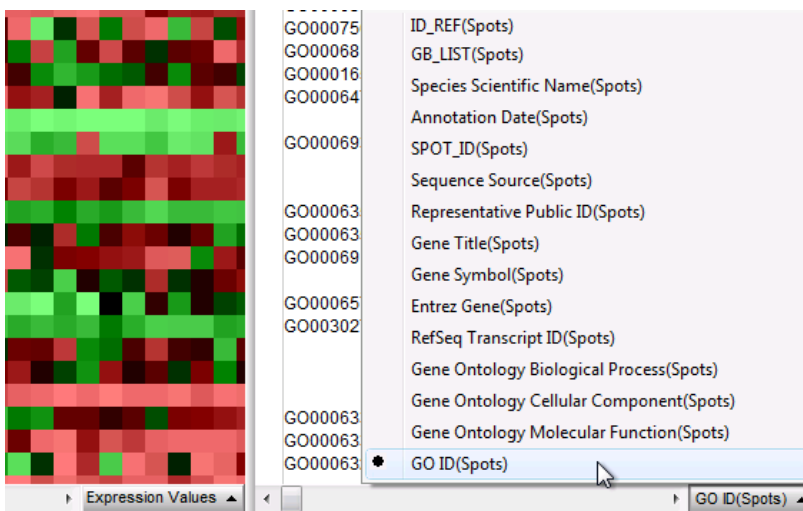


Figure 2-4. GO ID in the list of row identifiers.

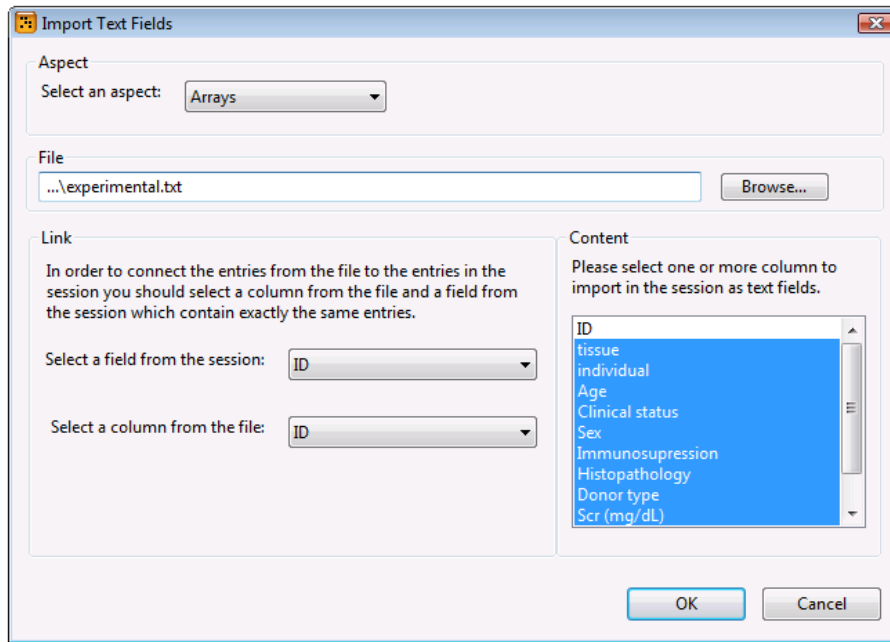


Figure 2-5. Adding information fields to the list of identifiers.

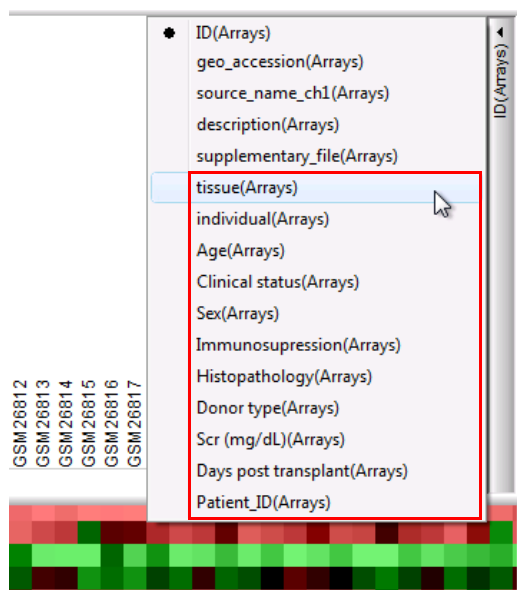


Figure 2-6. The new information fields in the list of column identifiers.

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 want to make row groups based on the information present in the row identifier GO ID (see Figure 2-5).

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

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

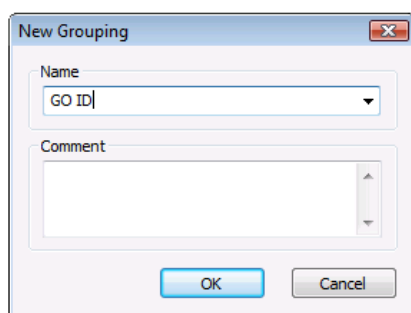


Figure 3-1. Grouping from the GO IDs.

3.1.3 GO ID is selected as the text field in the next window (see Figure 3-2). As you can see in the spot identifier list (see Figure 2-6), the same row can contain multiple GO IDs, separated by a “|” (e.g.

GO:0006955 | GO:0005529 | GO:0007166). Use a | (a pipe) as delimiter. This will split the multiple IDs for a certain row entry.

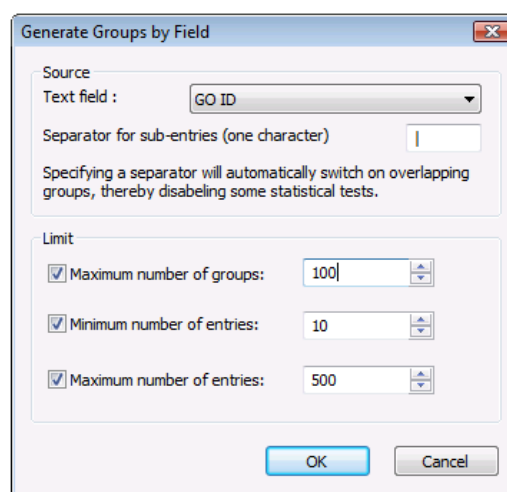


Figure 3-2. Creating groups based on the GO ID.

3.1.4 Change the settings in the *Limit* panel as specified in Figure 3-2 and click *<OK>*.

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

3.1.6 Press *<Exit>*.

In the *Row names* panel, the colors of the GO ID grouping are shown next to the row entry names (see Figure 3-4).

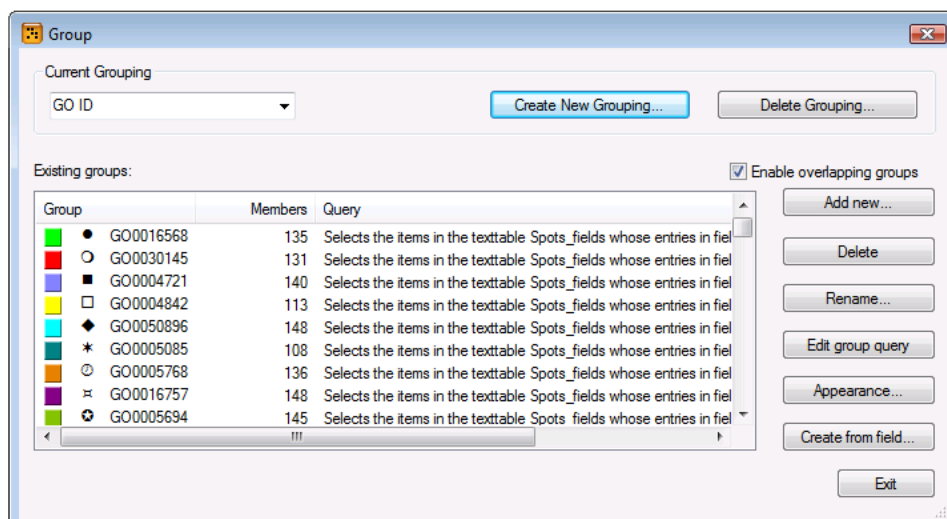


Figure 3-3. Groups based on GO IDs.

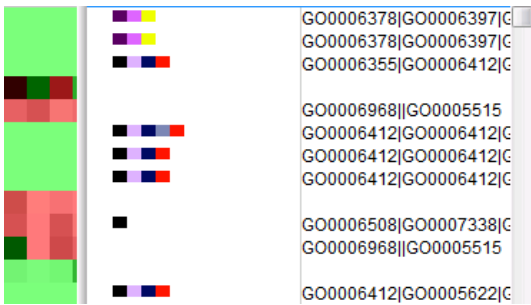


Figure 3-4. Colors based on the row groups.

GeneMaths XT offers the possibility to link a grouping to a website. In this example we will link the GO ID grouping to the GO-website. Later on we can use this link when using statistics reports.

3.1.7 Select *Groups > Row Group Link*.

3.1.8 In the next dialog box, select **GO ID** and http://www.godatabase.org/cgi-bin/amigo/go.cgi?action=replace_tree&query=### from the drop down lists. Click **<OK>** (see Figure 3-5).

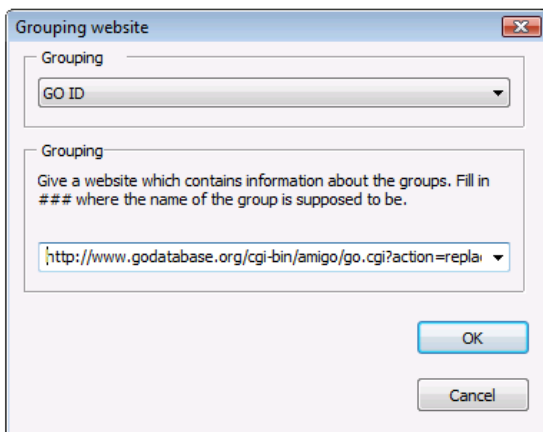


Figure 3-5. Linking a grouping to a website.

3.2 Column groups

In the next step, we want to make groups based on the column information fields tissue and individual.

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

3.2.2 In the next window, select **tissue** from the Name-pull down menu and click **<OK>** (see Figure 3-6).

3.2.3 Tissue is selected as the text field in the next window (see Figure 3-7). Uncheck ALL the limitations and click **<OK>**.

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

3.2.5 Repeat steps 3.2.1 until 3.2.3 for individual (see Figure 3-9).

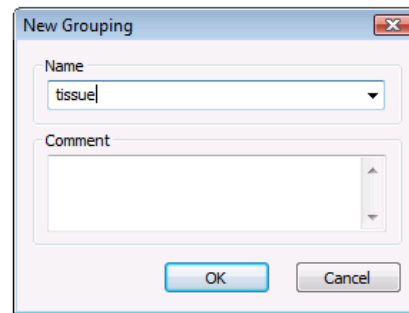


Figure 3-6. Grouping from tissue.

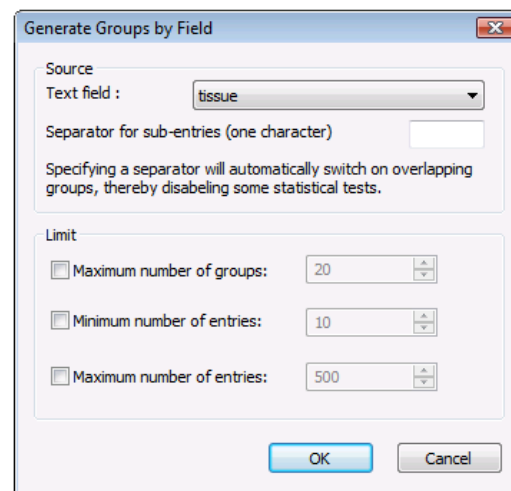


Figure 3-7. Creating groups based on the tissue.

3.2.6 The individual grouping consists of four groups (see Figure 3-9):

- Acute rejection
- Renal dysfunction w/o rejection
- Normal donor
- Well functioning transplant

3.2.7 Press **<Exit>**.

3.3 Group window

The drop down menu in the *Group* window lists all groupings defined in the database. When a grouping is selected from the drop down menu, the groups defined for that grouping are listed in the panel below.

3.3.1 Select for example the individual grouping from the drop down list.

All groups defined for this grouping are listed below the drop down list (see Figure 3-10).

Entries belonging to a group can be selected from within the *Group* window.

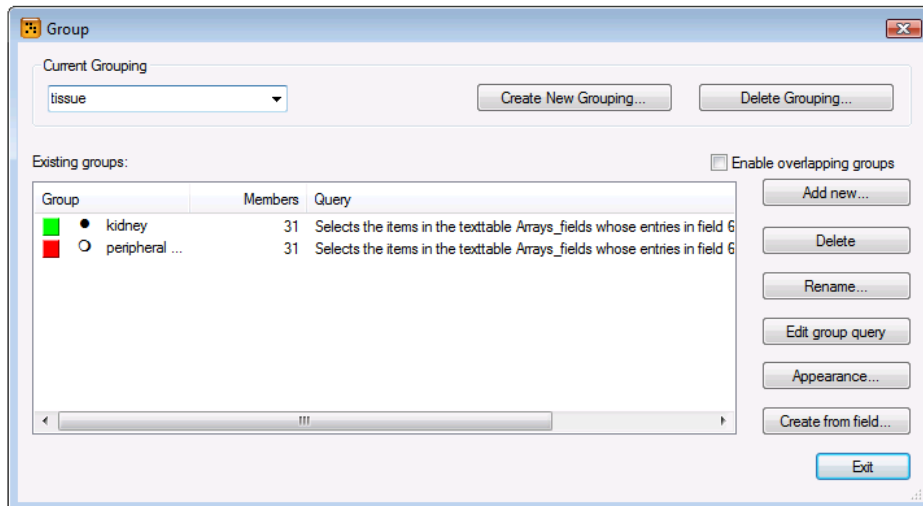


Figure 3-8. Groups based on the tissue.

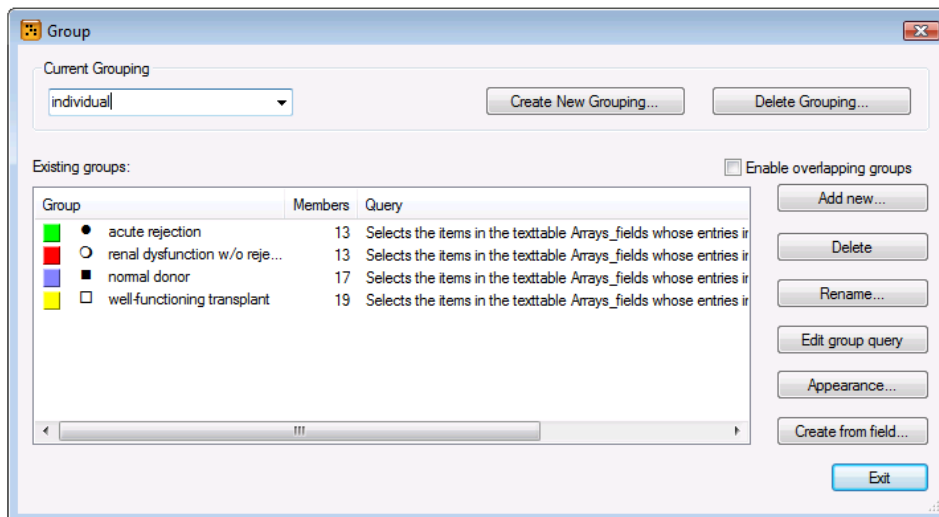


Figure 3-9. Groups based on 'individual'.

3.3.2 Hold the CTRL button and click on the square next to the group color. The square is highlighted in blue (see Figure 3-10) and the members of the group are selected in the data panel.

3.3.3 Use the SHIFT key to select more than one group at a time.

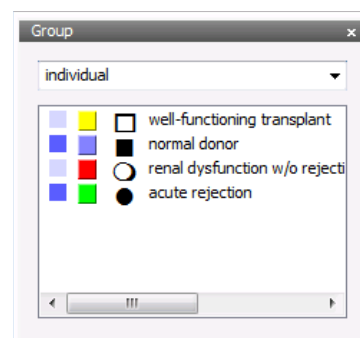


Figure 3-10. Group window.

4. Preprocessing

4.1 Log transformation

Before we can perform statistical tests we need transform our data, so we get a normal distribution. This is a prerequisite for these tests.

4.1.1 Select *Layer > Data Preparing > Log Transform* to calculate the log intensities of the layer **Target**. Store the results in a new layer called **LogTarget** (see Figure 4-1).

4.1.2 Press **<OK>**.

The log₂ intensities of the Target layer are calculated and stored in the LogTarget layer. This new layer is added to the *Layers* window.

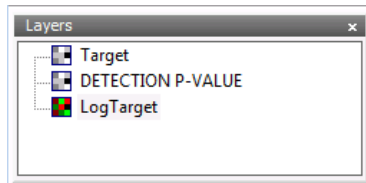


Figure 4-2. The new layer in the *Layers* window.

Make sure the **LogTarget** layer is selected in the *Layers* window. If you look at the histogram (top right panel), the shape is a Gaussian distribution. Note that the data is not yet centered around zero (see Figure 4-3).

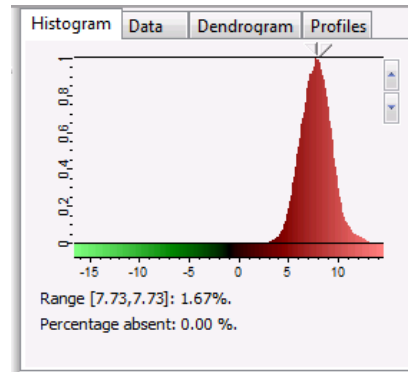


Figure 4-3. The 'LogTarget'-layer.

4.2 Analyze groups

In the next step we will take a closer look at the two tissue groups (peripheral blood lymphocyte and kidney tissue). It is likely to assume that both tissue groups actually can be seen as two different experiments.

4.2.1 Select *Profiles > Statistics Wizard*. Leave the orientation on **Row** and press **<Next>** (see Figure 4-4).

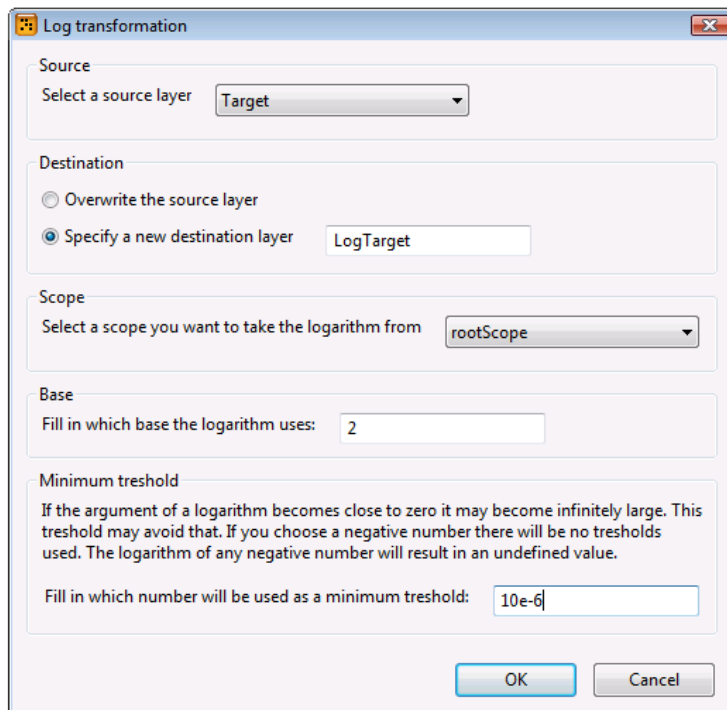


Figure 4-1. Settings for log transformation.

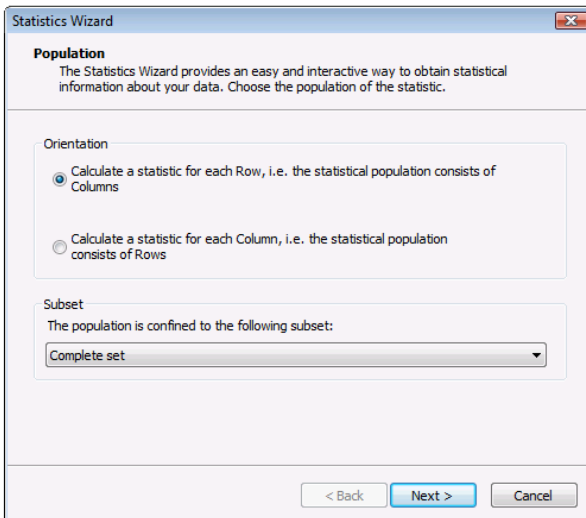


Figure 4-4. *Statistics Wizard: step 1.*

4.2.2 In the second step select the **Independent t-test** (under 'Independent test (two groups)') from the list and click <Next> (see Figure 4-5).

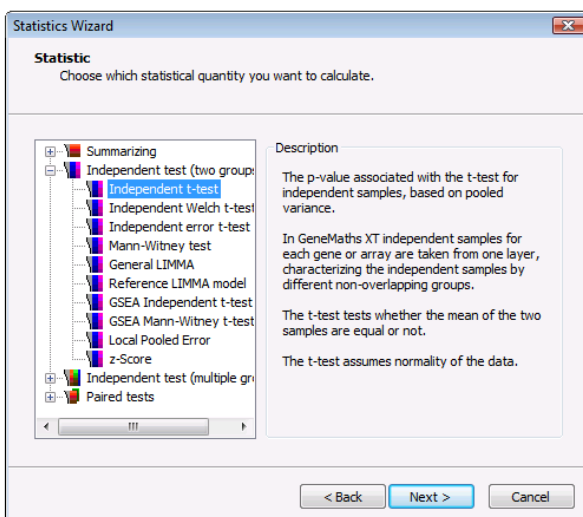


Figure 4-5. *Statistics Wizard: step 2.*

4.2.3 In the next window, make sure that **LogTarget** is selected and the two different tissue groups. Select **p-value** as output and click <Next> (see Figure 4-6).

4.2.4 Do not use a correction for multiple testing in the fourth step.

4.2.5 Click <Finish>.

4.2.6 Select the newly created profile in the *Profiles* tab. Click right on the profile name (see Figure 4-7) and select *Sort From Profile*.

4.2.7 Click right on the profile in the *Profile* panel and select *Show as Numbers* (see Figure 4-8).

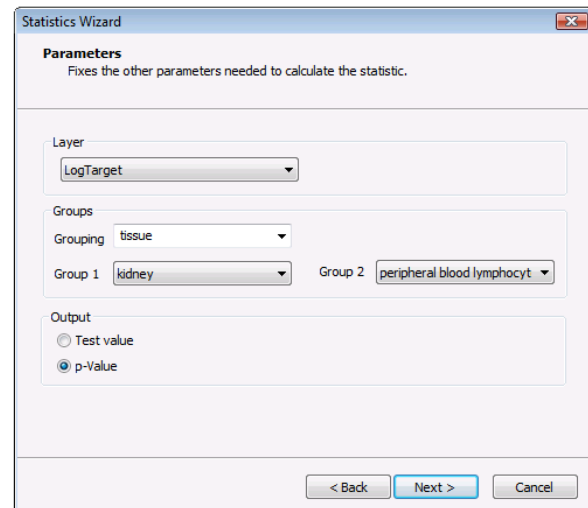


Figure 4-6. *Statistics Wizard, step 3.*

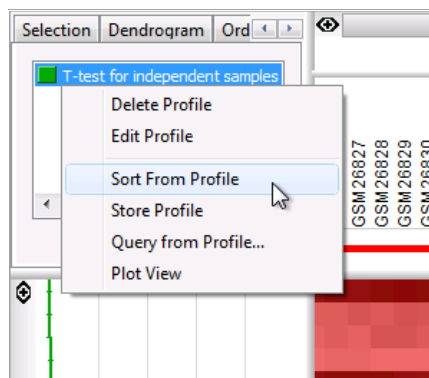


Figure 4-7. *Sort From Profile.*

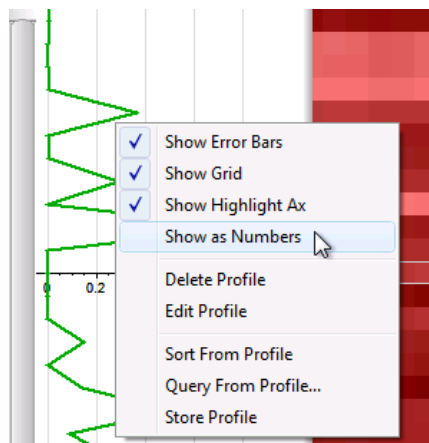


Figure 4-8. *Show as Numbers.*

In the *Profile* panel, the p-values for the row entries are shown. These p-values give an indication if the mean of the two groupings are the same. A lot of rows have a p-value close to 0 indicating that we have two experiments in the same dataset: the peripheral blood lymphocytes (PBL) and kidney tissue (BX).

The PBL and BX datasets will be treated as two experiments. We will create a subset for all the PBL arrays and another one for all the BX arrays and name them 'PBL' and 'BX' respectively.

4.2.8 Set **tissue** as the current grouping by clicking the down arrow next to the array groupings button



. Select **tissue** from the list.

4.2.9 To see the current grouping, select the **tissue** in the list of array identifiers (see Figure 2-6).

4.2.10 In the *Group* window, select **tissue** from the drop down menu. CTRL click on the square next to **tissue**.

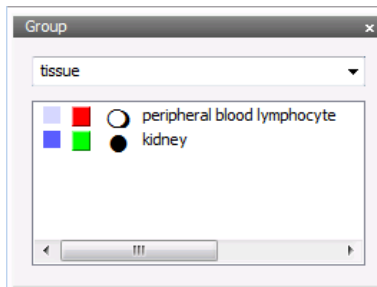


Figure 4-9. The *Group* window.

The specified arrays are selected in the *Column* panel and are indicated with a blue rectangle (see Figure 4-10).

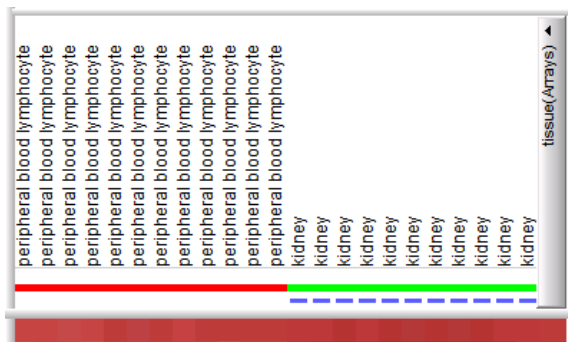


Figure 4-10. Selected arrays.

4.2.11 Select *Subset > Selection to Subset* and name the subset **BX**. Enable *Child of the complete set* and press <OK> (see Figure 4-11).

The selected arrays are stored in a subset of the complete subset (see Figure 4-12). The subset is added to the *Subsets* window.

4.2.12 Select the Complete set in the *Subsets* window.

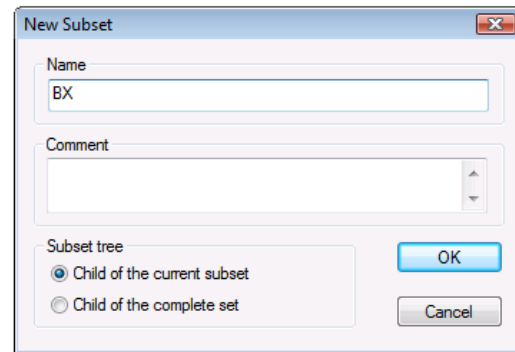


Figure 4-11. Create new subset.

4.2.13 Invert your selection with *Selection > Invert Column Selection*. All the PBL arrays are now selected, indicated with a blue rectangle.

4.2.14 Select *Subset > Selection to Subset* and name the subset **PBL**. Enable *Child of the complete set* and press <OK>.

In the *Main* window, the BX and PBL sets are both stored as a child of the 'Complete set' (see Figure 4-12). Clear the selection by pressing F4.

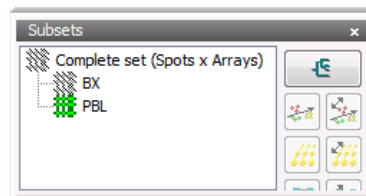


Figure 4-12. Complete set and the two subsets.

4.3 Filtering

In Figure 1-5, the fourth column represents the P/M/A-calls (Present/Missing/Absent). These calls are based on the p-values in the fifth column. These p-values are calculated from a one-sided Wilcoxon test and give an indication of the significance of the detection. The default cut-offs are "P" <= 0.04 < "M" <= 0.06 < "A". These values are present in the 'DETECTION P-VALUE'-layer.

We are going to clip our data values so that we have a new layer with only the present calls.

4.3.1 Select the **DETECTION P-VALUE** layer in the *Layers* window and the **Complete set** in the *Subsets* window.

4.3.2 Select *Layer > Filtering > Clip Extern*.

Fill out the dialog box as shown in Figure 4-13. With these settings, only the expression values of the **LogTarget** layer that have a value smaller than **0.04** (the P-calls) in the detection p-value layer are selected and stored in a layer called **Clipped**.

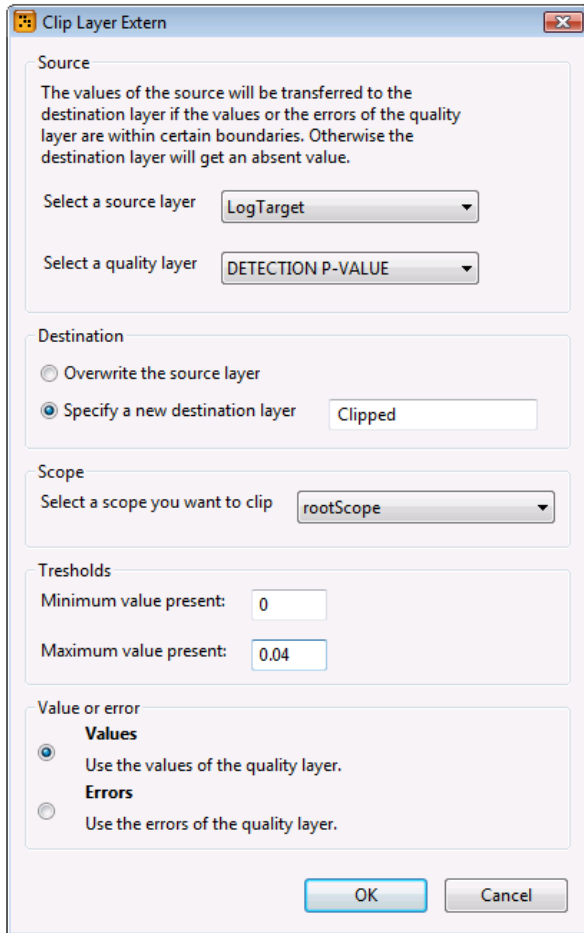


Figure 4-13. *Clip Values* dialog box.

4.3.3 Click <OK>.

As a first filter we will select all the row entries that have less than 100% absent calls. Genes without any expression value are useless anyway.

4.3.4 Select the **Clipped** layer and the **BX** subset in the *Main* window.

4.3.5 Select *Selection > Row Selection from Query* and click on the *Statistics Query* tab and then <*Statistics Builder*> (see Figure 4-14).

4.3.6 The *Statistics Wizard* opens. Make sure the orientation is set to **Row** and the **Complete** set is selected. Click <*Next*>.

4.3.7 In the second dialog, select the **Fraction absent values** statistic (under 'Summarizing') and click <*Next*> (see Figure 4-15).

4.3.8 In the next dialog, make sure you have the layer **Clipped** selected and click <*Next*> once more.

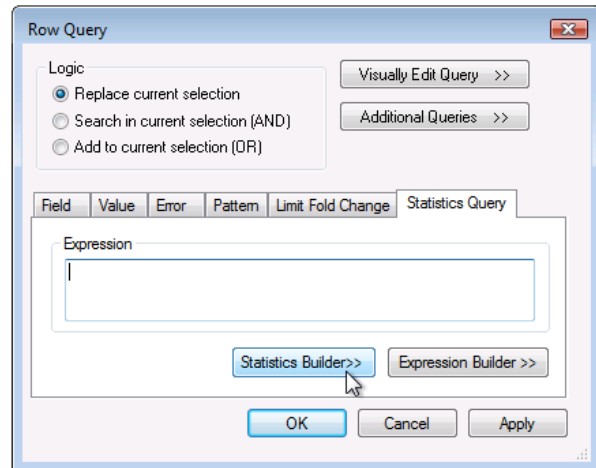


Figure 4-14. Row selection from query.

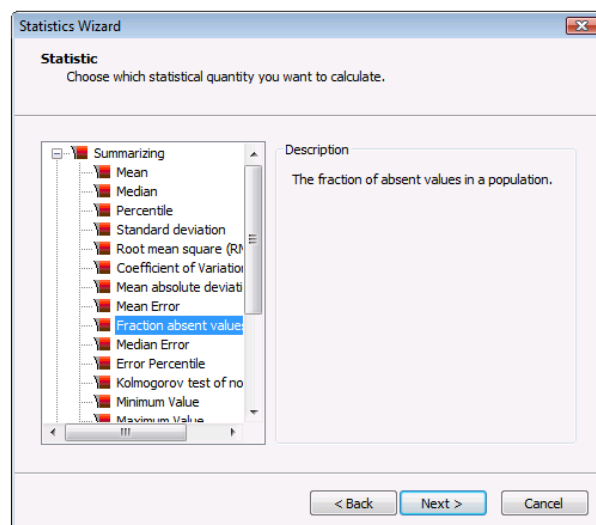


Figure 4-15. Choose which statistical quantity you want to calculate.

4.3.9 In the last window, set the value to < 1 (= 100%). GeneMaths XT will select all the row entries that have less than 100% absent calls, thus having at least one present call for at least one array. Click <*Finish*> and <*OK*> to close the query wizard.

The selected row entries (indicated with a blue arrow in the *Rows* panel) will be used for further analysis. First we need to store these selected row entries in a new subset. Make sure the subset **BX** is selected in the *Subsets* window.

4.3.10 Select *Subset > Selection to Subset* and give the new subset a name, e.g. **At Least 1 P-call in BX**. Store the selection in a new subset as a child of the current selected BX subset (Figure 4-16) and press <*OK*>.

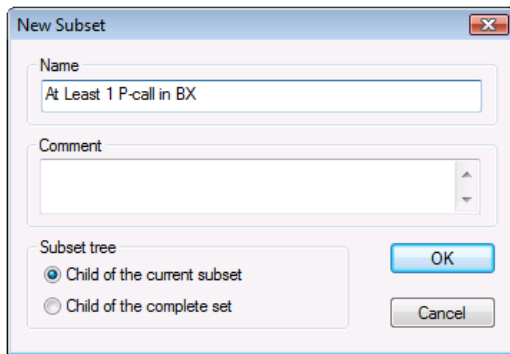


Figure 4-16. Create a new subset.

4.3.11 Clear the selection with F4

4.3.12 Repeat steps 4.3.4 - 4.3.9 for the PBL subset. Alternatively click CTRL+Q and change 'BX' in the *Expression* panel to 'PBL' (see Figure 4-17). Press <OK>.

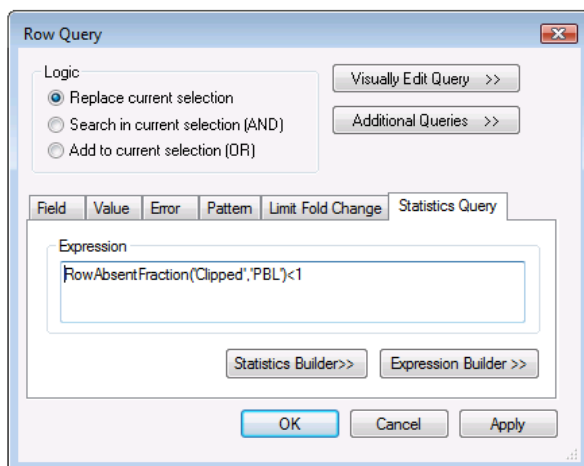



Figure 4-17. Row Query for PBL.

4.3.13 The selected files are indicated with a blue rectangle and will now be stored in a new subset. Select the 'PBL' subset in the *Subsets* panel in the *Main* window ().

4.3.14 Select *Subset > Selection to Subset* and give the new subset a name, e.g. 'At least 1 P-call in PBL'. Store the selection in a new subset as a child of the current PBL subset and press <OK>.

4.3.15 Clear the selection with F4. The *Subsets* panel should now look like this:

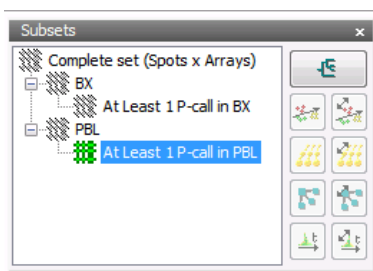


Figure 4-18. The *Subsets* window.

4.4 Normalization of arrays

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

4.4.1 Select *Layer > Normalization > Arrays*. Fill out the dialog box as shown in Figure 4-19. These settings will center our data around zero.

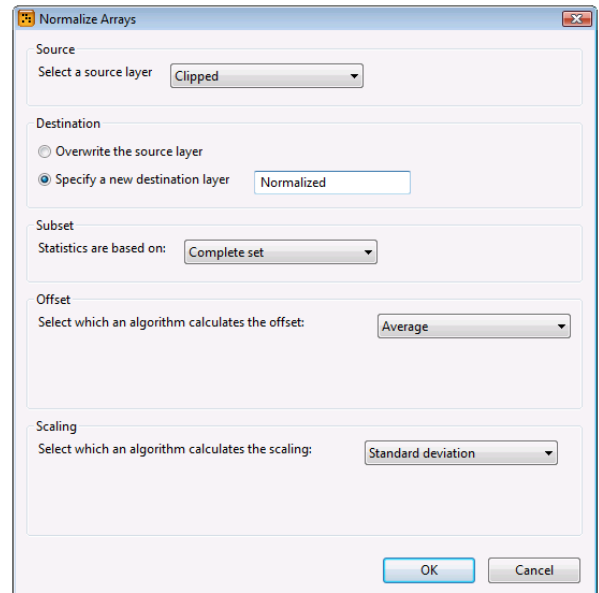


Figure 4-19. Settings for the normalization of the arrays.

4.4.2 Click <OK>.

The histogram of the new layer **Normalized** is now nicely centered around zero as shown in Figure 4-20.

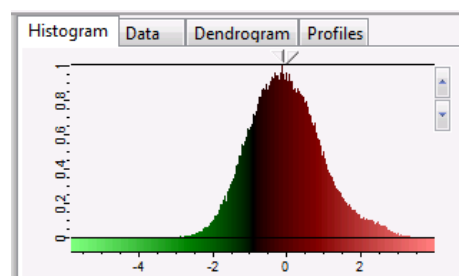


Figure 4-20. Data centered around zero.

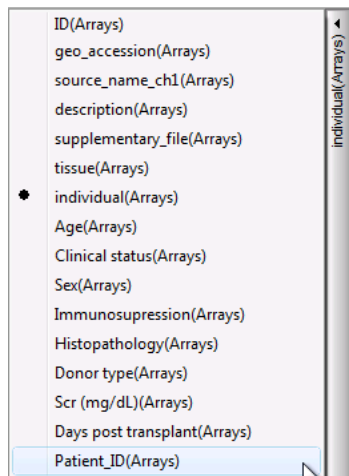
5. Statistics & Analysis

Now we have a starting point for further statistical analysis. Our aim is to find expression patterns that are unique for rejection, dysfunction without rejection and well functioning transplants for PBL and BX.

5.1 Hierarchical clustering


5.1.1 Make sure that no row/column entries are selected by pressing F4.

5.1.2 Select **Patient_ID** as the column identifier to address the arrays with these names.



- TX = well functioning transplant
- AR = acute rejection group
- C = normal donor
- NR = renal dysfunction w/o rejection

Click on the down arrow next to the array groupings

button  and select **individual**. Individual is now set as the current grouping (see Figure 5-1).

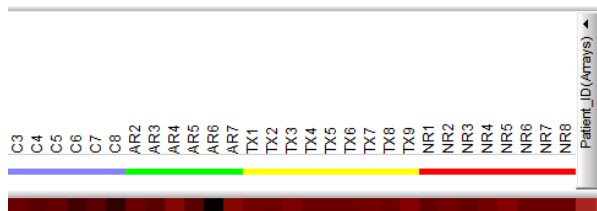
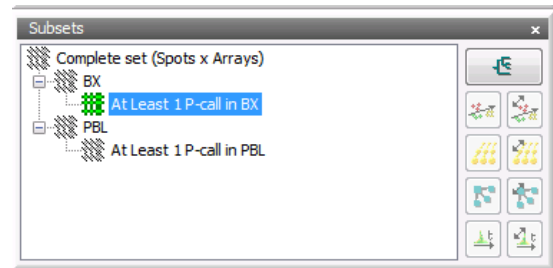



Figure 5-1. Colors correspond to the 'individual' grouping, the names to the Patient_ID information.

1) BX-subset

5.1.3 Select the **At Least 1 P-call in BX** subset in the *Subsets* window.



5.1.4 Select **Analysis > Cluster Analysis** or press . Fill out the dialog box as shown in Figure 5-2. Make sure the orientation is based on the **Columns** and press **<Next>**.

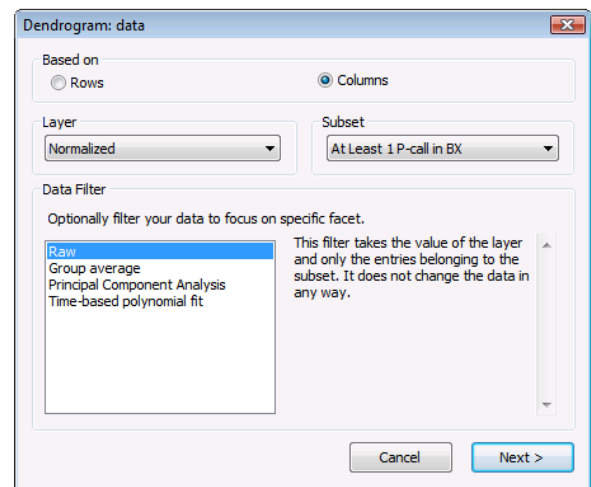


Figure 5-2. Cluster analysis: step 1.

We are going to create a UPGMA dendrogram using Pearson correlation as a similarity coefficient.

5.1.5 Make sure the **Pearson correlation** coefficient is selected and press **<Next>** (see Figure 5-3).

5.1.6 In the next step select **UPGMA** and press **<Finish>**.

5.1.7 GeneMaths XT will tell us that we have missing values in our layer. Because we are fully aware of this we will continue with these absent values. Press **<Yes>**.

5.1.8 The dendrogram pops up in the *Main* window.

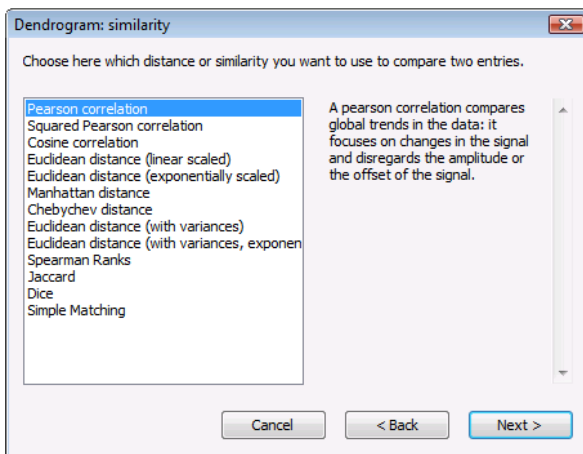


Figure 5-3. Settings for the cluster analysis.

5.1.9 You can select branches by pressing CTRL and the left mouse button while pointing to the branch node. The arrays belonging to that branch are selected. Pressing F4 clears the selection.

5.1.10 Select a branch and click on the right mouse button. A menu pops up. Select *Swap branch*. The members of the selected branch are swapped. Please note that the swapping branches does not alter the similarities in a dendrogram, it is just a different way of visualization.

5.1.11 Swapping the branches of the BX subset gives an indication of the clustering (see Figure 5-4). The members of the 'individual' grouping cluster relatively good together, except for the 'renal dysfunction w/o rejection group' (= NR). Because this can be a disturbing factor in further analysis, we will create a new subset with only the arrays that cluster well together and exclude the 'renal dysfunction w/o rejection' group.

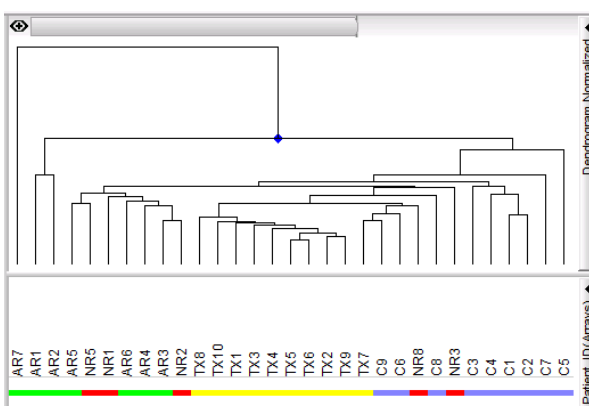


Figure 5-4. Dendrogram of the BX subset.

5.1.12 In the *Group* window, select **individual** from the list. Select all groups except for the 'renal dysfunction w/o rejection' group by CTRL clicking on the square next to the group.

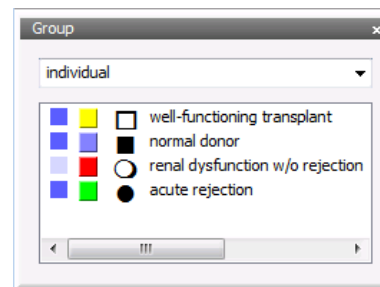



Figure 5-5. The *Group* window.

5.1.13 Select *Subset > Selection to Subset* and store the selection as **No NR**. Enable *Child of the current subset* and press <OK>.

Next, we are going to make a dendrogram for this new subset.

5.1.14 Select *Analysis > Cluster Analysis* or press . Make sure the orientation is based on the **Columns** and the **No NR** subset and the **Normalized** layer are selected. Press <Next> twice and then <Finish>.

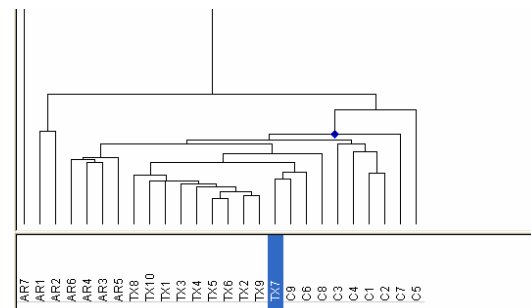


Figure 5-6. Dendrogram of a selection of the BX subset.

2) PBL-subset

5.1.15 Press F4 to unselect the arrays.

5.1.16 Repeat steps 5.1.3. to 5.1.8 for the **At Least 1 P-call in PBL** subset.

5.1.17 Swapping the branches of the PBL subset gives an indication of the clustering. The 'Acute Rejection' group (= AR) does not cluster well (see Figure 5-7).

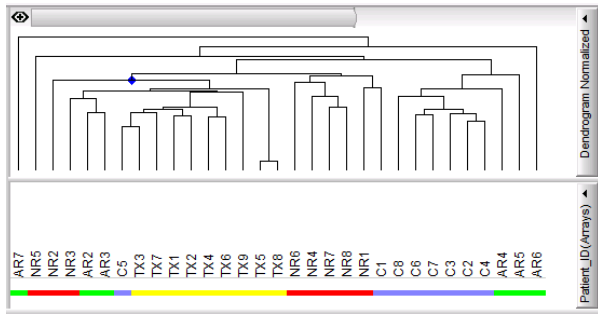



Figure 5-7. Dendrogram of PBL subset.

5.1.18 In the *Group* window, select all groups of the **individual** grouping, except for the 'Acute Rejection' group.

5.1.19 Select *Subset > Selection to Subset* and store the selection as **No AR**. Enable *Child of the current subset* and press <OK>.

5.1.20 Make a dendrogram for this new subset. Select *Analysis > Cluster Analysis* or press . Make sure the orientation is based on the **Columns** and the **No AR** subset and the **Normalized** layer are selected. Press <Next> twice and then <Finish>.

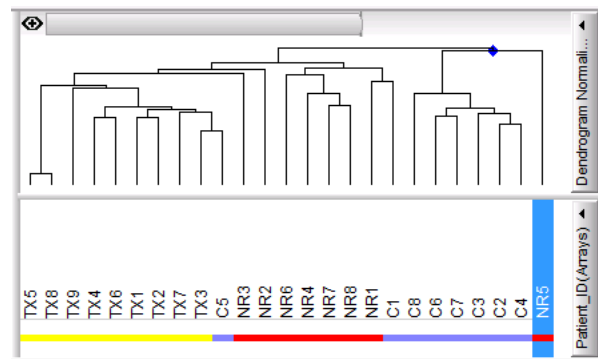


Figure 5-8. Dendrogram of a selection of the BX subset.

5.1.21 Swap the branch in such a way that the dendrogram looks like the one in Figure 5-8. Press F4 to unselect all arrays and select the TX-branch (CTRL + left click on the branch node) All arrays belonging to this branch are selected (see Figure 5-9).

Statistics report

Grouping: individual
 Subset: No AR
 Total number of items: 25
 Selected number of items: 10
 Expected ratio: 0.4

Category	Binomial p-Value	Hypergeometric p-Value	Total	Partial	Ratio	z-Score
well-functioning transplant	6.79535e-006	4.89482e-006	9	9	1	4.5
renal dysfunction w/o rejection	0.00606997	0.00594966	8	0	0	-2.74398
normal donor	0.0592297	0.0654462	8	1	0.125	-1.88648

Figure 5-10. Statistics report.

5.1.22 Right click on the branch and select *Branch to Statistics Report* (see Figure 5-9).

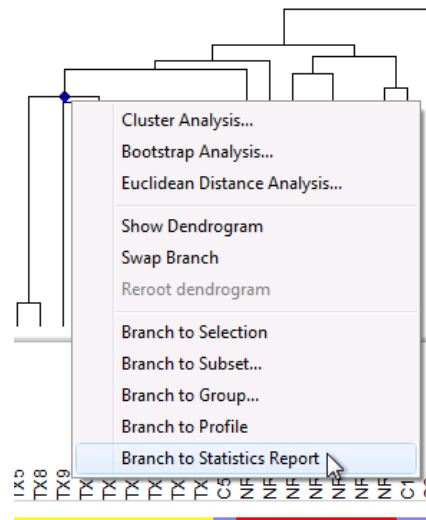


Figure 5-9. Opening the statistics report of the selected branch.

5.1.23 The report gives the odds of having the C5 in the TX branch (see Figure 5-10).

The array with Patient_ID C5 (normal donor) has 40% (= 0.4) chance of being present in the TX branch. The actual ratio of being present in the TX branch is 12.5% (= 0.125), this corresponds to a p-value of about 0.06. This means that the odds of having C5 in the TX branch is much less than by chance.

3) Remarks

- Please note that a dendrogram based on the 'LogTarget'-layer will look exactly the same as the one based on the 'Normalized'-layer when using Pearson. Why?
- What if you use Euclidean distance instead of Pearson, does this influence the shape of the dendrogram in both layers?
- If we would use the median/MAD instead of the mean/stdev as parameters for the normalization of the arrays, would this influence the results? Why (not)?

5.2 Differentially expressed genes

In this part we will try to find the differentially expressed row entries for each clustered group.

1) Kidney tissue (BX)

For the kidney tissue (BX) subset we will compare:

- Well functioning transplant (TX) versus an acute rejection (AR)
- Well functioning transplant (TX) versus a normal donor (C)

TX vs. AR

First we are going to select all differentially expressed genes with a T-test. We will use a 0.001 significance level and correct for multiple testing. Make sure that no row or column entries are selected by pressing F4.

5.2.1 Select *Selection > Row Selection from Query* (or press CTRL+Q) to open the *Row Query* dialog box. Click on the *Statistics Query* tab and then press on *<Statistics Builder>*.

5.2.2 Select the **No NR** subset and click *<Next>*.

5.2.3 In the next window, select the **Independent t-test** (under 'Independent test (two groups)') and click *<Next>* (see Figure 5-11).

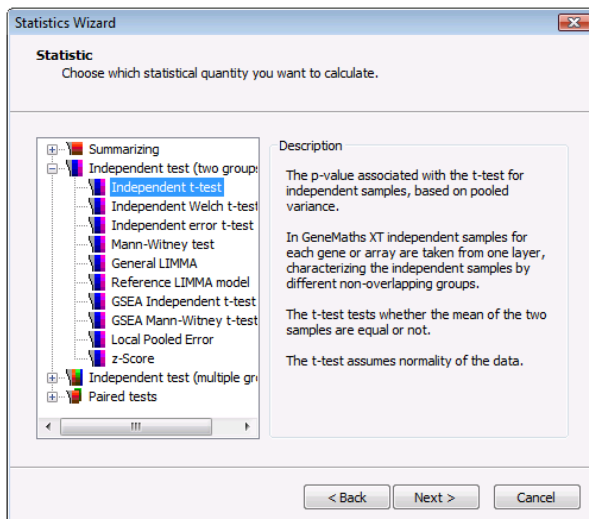


Figure 5-11. Choose statistical quantity.

5.2.4 Fill out the settings in the next window as shown in Figure 5-12.

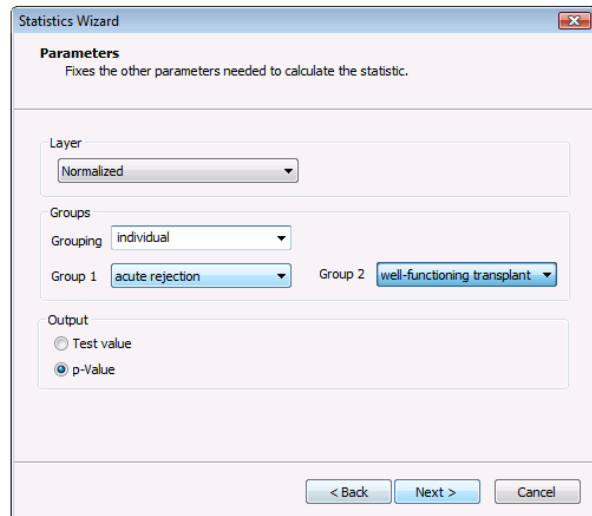


Figure 5-12. Parameters to calculate the statistic.

5.2.5 In the next step, choose *Benjamini & Hochberg procedure* to control the false discovery rate (FDR). Press *<Next>*.

5.2.6 In the final step, set the threshold to < 0.001 (see Figure 5-13). Click *<Finish>* and then press *<OK>*.

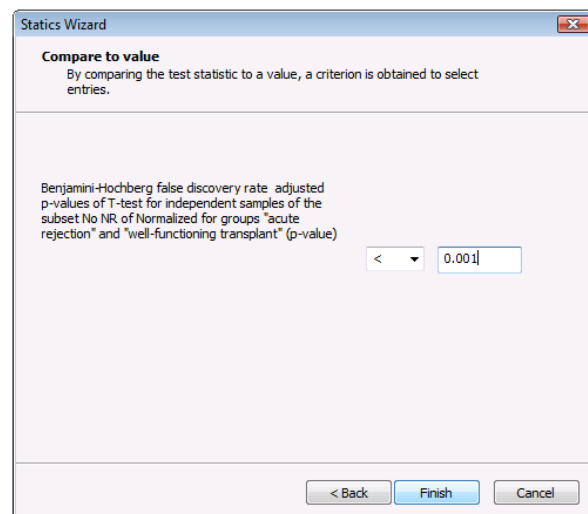


Figure 5-13. Setting a criterion for the test statistic.

In a next step we are going to store the selected row entries.

5.2.7 Select *Selection > Store Selection* and store the **Row** selection as a new query called **T-test**. Press **<OK>**.

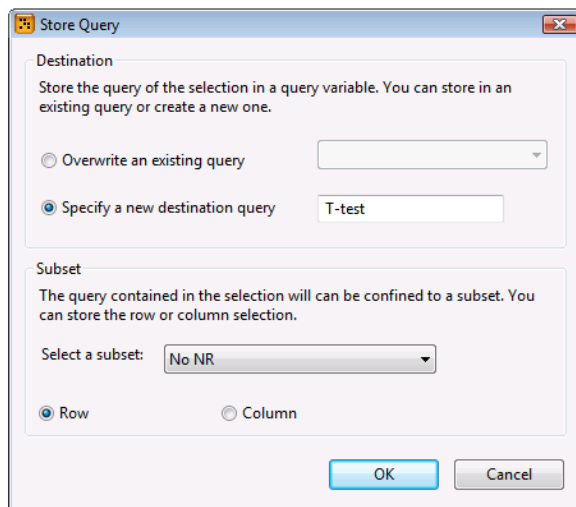


Figure 5-14. Store the row query.

The genes are now selected, now we want to select the TX (Well functioning transplant) and AR (acute rejection) arrays.

5.2.8 In the *Group* window, make sure the **individual** grouping is selected from the pull down list. Select the TX (Well functioning transplant) and AR (acute rejection) groups by CTRL clicking on the square next to their name (see Figure 5-15).

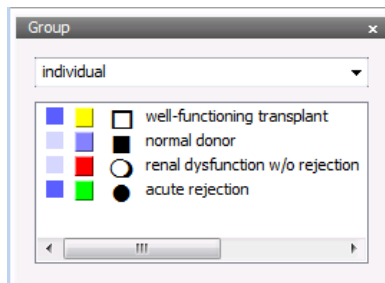


Figure 5-15. The *Group* window.

In the *Main window*, the TX and AR arrays are now selected in addition to the previous selected row entries. We are now going to create a new subset called **AR vs. TX** for our selection.

5.2.9 Select the No NR subset in the *Subset* panel.

5.2.10 Select *Subset > Selection to Subset* and give the new subset the name **AR vs. TX**. Store the selection in a new subset as a child of the current No NR subset and press **<OK>**.

Next, we are going to select all the row entries that have less than 1% absent calls, thus having at least 99% present calls.

5.2.11 Select the AR vs. TX subset in the *Subsets* window.

5.2.12 Select *Selection > Row Selection from Query* and click on the *Statistics Query* tab and then **<Statistics Builder>**.

5.2.13 The *Statistics Wizard* opens. Leave the first dialog unaltered and click **<Next>**.

5.2.14 Select the **Fraction absent values**-statistic (under 'Summarizing') and click **<Next>**.

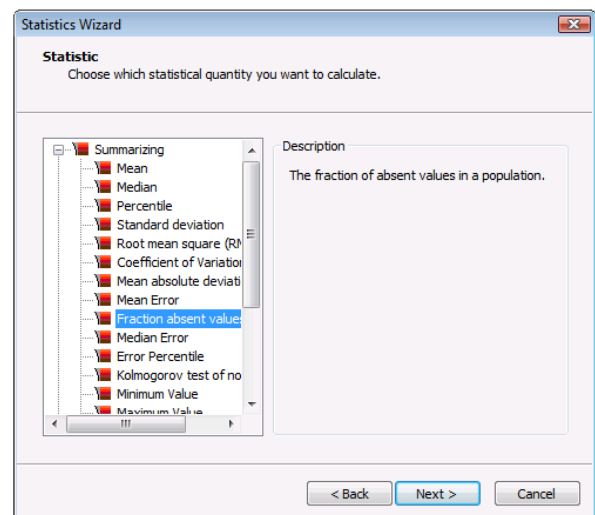


Figure 5-16. *Statistics wizard: step 2.*

5.2.15 In the next dialog, make sure you have the layer **Normalized** selected and press **<Next>**.

5.2.16 In the last window, set the value to **< 0.01**. Click **<Finish>** and **<OK>** to close the query.

Next, we are going to create a new subset for the selected genes.

5.2.17 Select the AR vs. TX subset in the *Subset* panel.

5.2.18 Select *Subset > Selection to Subset* and give the new subset the name **Present**. Store the selection in a new subset as a child of the current AR vs. TX subset and press **<OK>**.

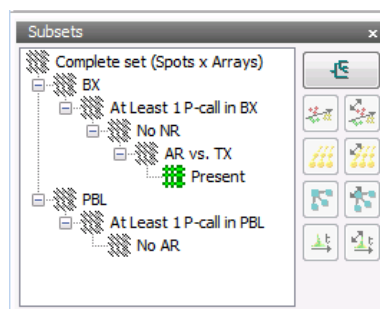


Figure 5-17. The *Subsets* window.

5.2.19 Normalize the row entries with *Layer* > *Normalization* > *Genes* and store the result in a new layer 'NormGenes TX_AR' (see Figure 5-18).

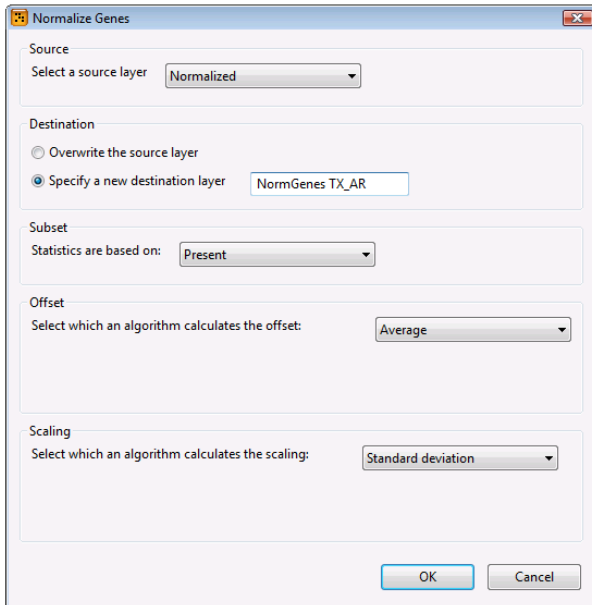





Figure 5-18. Normalize the genes.

Next, we are going to create a dendrogram for both the column and row entries.

5.2.20 Press the  button. Make sure the **Present** subset and the **NormGenes TX_AR** are selected. Press next twice and press <Finish>.

5.2.21 Press the  button. Make sure the **Present** subset and the **NormGenes TX_AR** are selected. Press next twice and press <Finish>.

5.2.22 To obtain the optimal view of your dendrograms, use the zoom buttons () or the zoom slider (see Figure 5-19).

5.2.23 Unselect the current selection by pressing F4.

5.2.24 Select the AR branch and the upregulated genes (red color) for this branch (see Figure 5-19): 10 genes are upregulated in AR (red color) and 128 genes (green color) are downregulated in AR.

Branches can be abridged to see the average values of the branch.

5.2.25 Select the branch with the 128 gene members by clicking on the node. Right click and select *Abridge Branch* (see Figure 5-20).

5.2.26 Select the branch with the 10 overexpressed genes for AR by clicking on the node. Select *Selection* > *Row Selection to Statistics Report*.

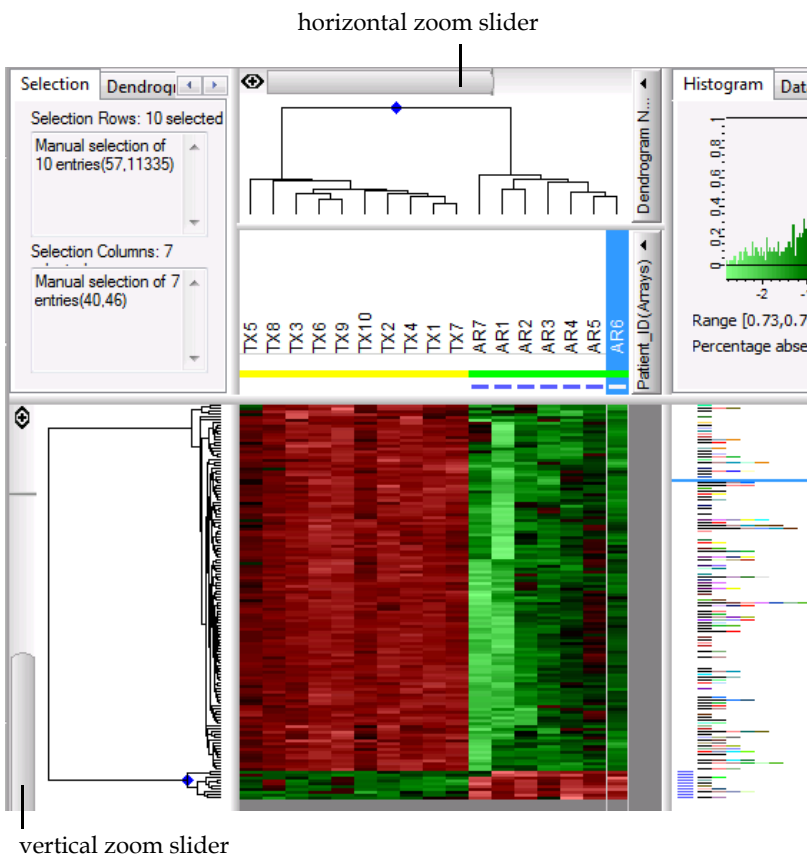


Figure 5-19. Two fold clustering.

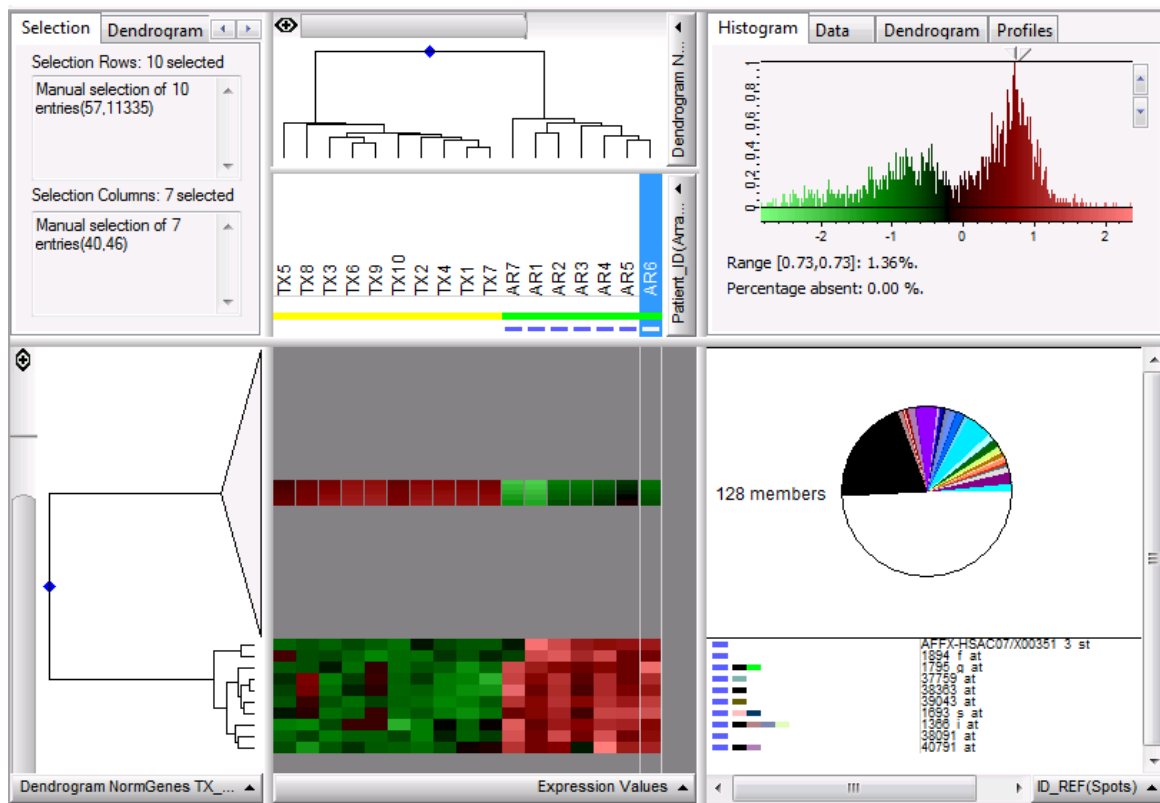


Figure 5-20. Abridge branches to obtain a better view.

This report (see Figure 5-21) shows which GO ID is present more or less than expected based on an odd ratio.

Click on one of the hyperlinks. The GO website opens, showing the ontology of this term. Based on such report it is possible to see which GO terms are over/under expressed in the selected branch.

TX vs. C

- Try now to find the up/down regulated genes for TX vs. C.
- Are the same genes up/down regulated?

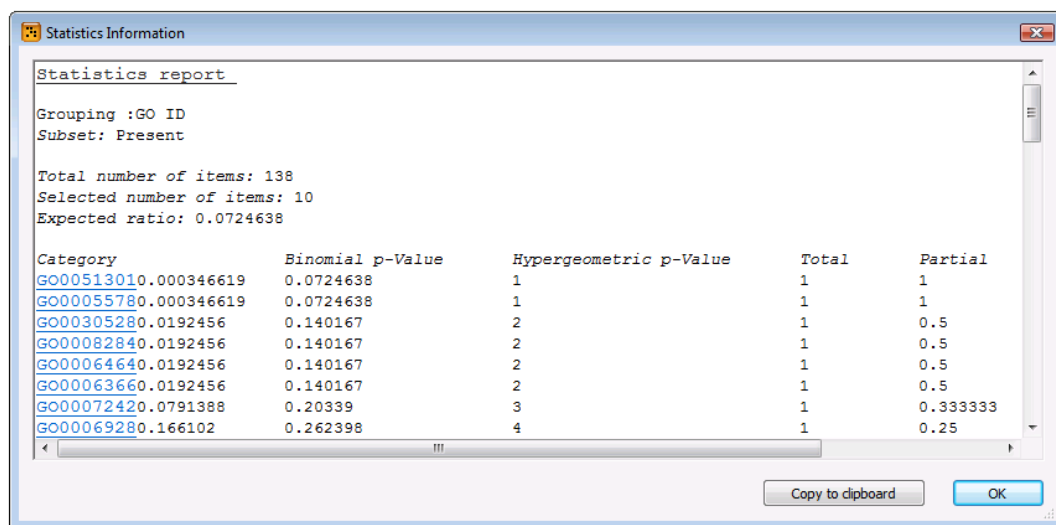


Figure 5-21. Statistics report.

2) Peripheral blood lymphocyte (PBL)

- Repeat the analysis above but now for the PBL dataset.
- Find up/down regulated genes for NR vs. TX and TX vs. C.
- Are the same genes up/down regulated?
- Is there a relation between the genes found in TX vs. C for the BX dataset and the ones found in the TX vs. C subset for PBL?

5.3 Comparing statistical tests

Differentially expressed genes can be screened for by several statistical tests. In the previous steps (5.2.1 - 5.2.6) we detected differentially expressed genes in the NoNR subset between TX and AR with an independent T-test (threshold of <0.001 and a correction for multiple testing). These differentially expressed genes were stored as **T-test** (see 5.2.7).

Next we are going to perform two additional tests on this subset and look if the three tests denote more or less the same genes as 'differentially expressed'.

5.3.1 Select the **No NR** subset in the *Subsets* window.

5.3.2 Select *Selection > Row Selection from Query* (or press CTRL+Q) to open the *Row Query* dialog box. Click on the *Statistics Query* tab and then **<Statistics Builder>**.

5.3.3 Select the **No NR** subset and click **<Next>**.

5.3.4 In the next window, select the **Local Pooled Error** (under 'Independent test (two groups)') and click **<Next>**.

5.3.5 Fill out the settings in the next window as shown in Figure 5-12.

5.3.6 In the next step, choose *Benjamini & Hochberg procedure* to control the false discovery rate (FDR) and click **<Next>**.

5.3.7 In the final step, set the threshold to < 0.001. Click **<Finish>** and then press **<OK>**.

5.3.8 Store the selected row entries with *Selection > Store Selection* and name the selection **LPE**. Make sure the orientation is set to **Row** and the **No NR** subset is selected. Press **<OK>**.

5.3.9 Repeat step 5.3.2 - 5.3.7 but now choose the **Reference LIMMA model** (under 'Independent test (two groups)') as test statistic.

5.3.10 Store the selected genes with *Selection > Store Selection* and name the selection **LIMMA test**.

Next we are going to look if the three tests (LIMMA test, LPE and T-test) denote more or less the same row entries as 'differentially expressed'.

5.3.11 Select *Selection > Venn Diagram*. Select the three stored queries and press **<OK>**.

The Venn diagram pops up in a new window.

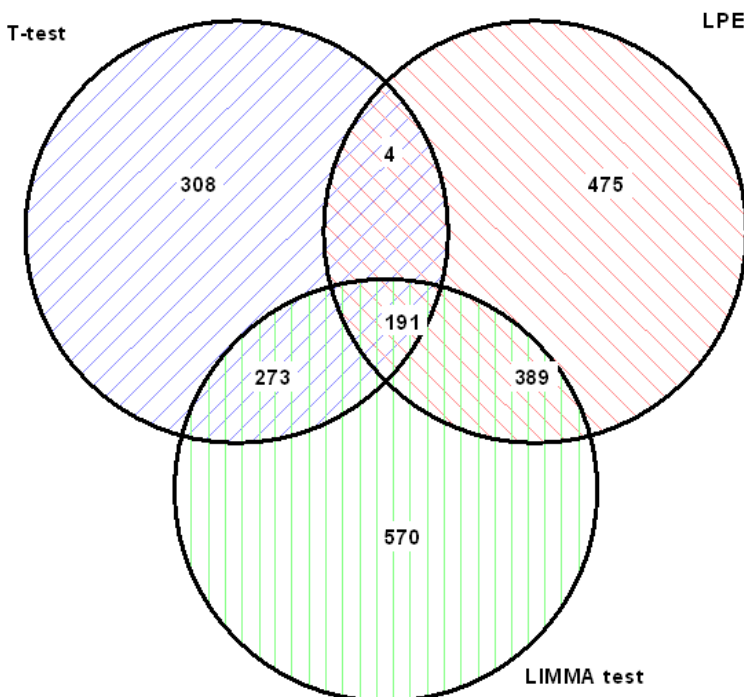


Figure 5-22. Venn Diagram.