

DNA Microarray Data Analysis

IIRIS HOVATTA, KATJA KIMPPA, ANTTI LEHMUSSOLA, TOMI PASANEN,
JANNA SAARELA, ILANA SAARIKKO, JUHA SAHARINEN, PEKKA TIIKKAINEN
TEEMU TOIVANEN, MARTTI TOLVANEN, MAUNO VIHINEN AND GARRY WONG
EDITORS JARNO TUIMALA AND M. MINNA LAINE

CSC

CSC – Scientific Computing Ltd. is a non-profit organization for high-performance computing and networking in Finland. CSC is owned by the Ministry of Education. CSC runs a national large-scale facility for computational science and engineering and supports the university and research community. CSC is also responsible for the operations of the Finnish University and Research Network (FUNET).

All rights reserved. The PDF version of this book or parts of it can be used in Finnish universities as course material, provided that this copyright notice is included. However, this publication may not be sold or included as part of other publications without permission of the publisher.

© The authors and
CSC – Scientific Computing Ltd.
2005

Second edition

ISBN 952-5520-11-0 (print)

ISBN 952-5520-12-9 (PDF)

<http://www.csc.fi/oppaat/siru/>

<http://www.csc.fi/molbio/arraybook/>

Printed at
Picaset Oy
Helsinki 2005

2 Web extra: Basic statistics

2.1 Statistics using GeneSpring

A few examples of statistical data manipulation, simple statistics calculation and statistical testing using the DNA microarray data analysis software GeneSpring are given in this section.

2.1.1 Simple statistics

Genewise Simple statistics, like the average, minimum, maximum, standard error of the mean, and standard error can be produced through *Edit->Copy->Copy Annotated Genelist* in GeneSpring. After pasting the information in Excel or other spreadsheet program, the values will become apparent. These simple statistics can be produced for intensities of either channel (raw and control data in GeneSpring), intensity ratio and log-transformed intensity ratio (normalized data in GeneSpring).

2.1.2 Transformations

In GeneSpring, data transformations are linked to Experiment Interpretation. There are three options to choose from: Non-transformed data (ratio) \log_2 -transformed data (log of ratio) and fold change . When one transformation is chosen, GeneSpring will automatically recalculate the data values, and use the new values for any subsequent statistical analyses (statistical group comparison, k-means clustering, etc.).

2.1.3 Scatter plot and histogram

A scatter plot can be produced in GeneSpring through *View->Scatter plot*. The shown axes can be modified from *View->Display Options*. A scatter plot can easily be used for testing the linearity of the data.

A histogram is displayed, if you have selected the *View->View Graph*, and additionally have set up parameters so that the quantification results are shown separately for every hybridized chip. For example, in a simple time series experiment, setting time as a non-continuous parameter would produce a histogram of expression values. You can use histograms for assessing the distribution of the data. After log-transformation the distribution of expres-

sion values should be symmetric and one-peaked.

2.1.4 Correlation

The Pearson correlation between chips is automatically calculated. The values of correlation coefficients can be viewed through Condition Inspector. Condition Inspector is invoked when the right-hand mouse button is clicked over one chip in the navigator bar, and Inspect is selected. From the opening window, select the Similar Conditions tab. The correlation coefficients between the selected chip and all the other chips are reported in the column Correlation (Figure 2.1).

Correlation	Experiment Name	Mouse	Signal channel
0.66275	Mouse testis again	2	3
0.59187	Mouse testis again	6	3
0.58806	Mouse testis again	5	3
0.56766	Mouse testis again	4	3
0.48144	Mouse testis again	3	3
-0.31119	Mouse testis again	4	5
-0.34872	Mouse testis again	3	5
-0.44185	Mouse testis again	5	5
-0.48318	Mouse testis again	6	5
-0.49606	Mouse testis again	2	5
-0.53548	Mouse testis again	1	5

Figure 2.1: The Pearson correlation in GeneSpring is found under the Similar Condition-tab in Condition Inspector.

2.1.5 Linear regression

GeneSpring can calculate a linear regression model producing a line of best fit for a 2D scatter plot view. The line of best fit is produced from *View->Display options*. Select the Lines to Graph tab, and tick Line of Best Fit box. The linear regression line is overlaid with the scatter plot, and the regression equation of form $y = aX + b$ is displayed at the bottom of the scatter plot view. Recall that a in the regression equation equals the correlation coefficient between the two variables plotted along the axes.

2.1.6 One-sample t-test

The one sample t-test in GeneSpring is automatically calculated for all the genes, whenever replicates are available. The t-test p -values can be found from the Gene Inspector, which is opened by double-clicking the left mouse button over a gene (Figure 2.2). The same p -values are also reported in the Spreadsheet, which is invoked from *File->View as spreadsheet*.

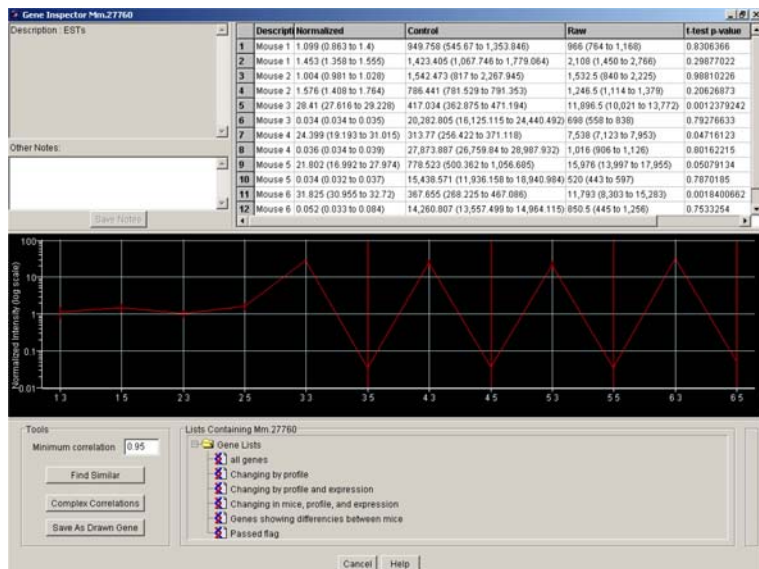


Figure 2.2: In GeneSpring the p -values for the one-sample t-test are found from Gene Inspector

2.1.7 Independent samples t-test and ANOVA

The independent samples t-test and ANOVA are located in *Tools->Filtering and statistical analysis*. Clicking the right hand mouse button on an experiment in the opening window enables one to Add Statistical Group Comparison. You can specify the parameters by which the compared groups are defined, select the groups to compare, select the appropriate statistical test and adjust p -value and multiple testing correction (Figure 2.3).

The result of the statistical group comparison is a list of genes, which are statistically differentially expressed between the specified groups. The actual p -values for these genes can be found from the Gene List Inspector, which can be opened by clicking a gene list with the right hand mouse button, and selecting Inspect (Figure 2.4).



Figure 2.3: In GeneSpring statistical group comparison is a tool of its own.

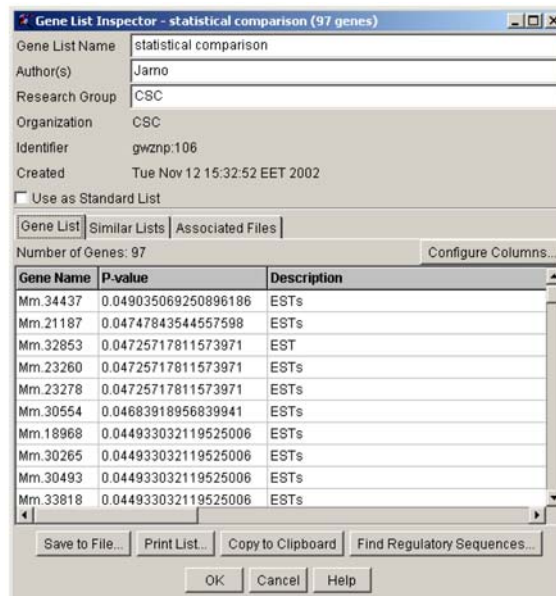


Figure 2.4: The result of the statistical group comparison is stored into a genelist, which can be viewed with a Genelist Inspector.