



## 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

## List of Contributors

Iiris Hovatta  
National Public Health Institute  
Haartmaninkatu 8  
FI-00290 Helsinki  
Finland

Katja Kimppa  
PerkinElmer Life Sciences and Analytical Sciences  
- Wallac Oy  
P.O.Box 10  
FI-20101 Turku  
Finland

M. Minna Laine  
CSC, the Finnish IT center for science  
Keilaranta 14  
FI-02101 Espoo  
Finland

Antti Lehmussola  
Tampere University of Technology  
P.O.Box 553  
FI-33101 Tampere  
Finland

Tomi Pasanen  
University of Helsinki  
P.O.Box 68  
FI-00014 University of Helsinki  
Finland

Janna Saarela  
Biomedicum Biochip Center  
Haartmaninkatu 8  
FI-00290 Helsinki  
Finland

Ilana Saarikko  
University of Helsinki  
P.O.Box 68  
FI-00014 University of Helsinki  
Finland

Juha Saharinen  
National Public Health Institute  
Haartmaninkatu 8  
FI-00290 Helsinki  
Finland

Pekka Tiikkainen  
VTT  
P.O.Box 106  
FI-20521 Turku  
Finland

Teemu Toivanen  
Centre for Biotechnology  
Tykistökatu 6  
FI-20521 Turku  
Finland

Martti Tolvanen  
Institute of Medical Technology  
Biokatu 8  
FI-33520 Tampere  
Finland

Jarno Tuimala  
CSC, the Finnish IT center for science  
Keilaranta 14  
FI-02101 Espoo  
Finland

Mauno Vihinen  
Institute of Medical Technology  
Biokatu 8  
FI-33520 Tampere  
Finland

Garry Wong  
A. I. Virtanen -institute  
University of Kuopio  
FI-70211 Kuopio  
Finland

# 1 Web extra: Preprocessing of data

Jarno Tuimala

## 1.1 Examples using GeneSpring

This section describes some preprocessing examples using the GeneSpring program.

### 1.1.1 Importing data

Before any preprocessing can be done for the data using GeneSpring, it should be imported into the program. GeneSpring read virtually any file format if it is a tab-delimited text file. In GeneSpring you need to specify what kind of information some key columns contain. Basically, the minimal datafile contains just two columns: one for the gene name (it can also be Genbank accession number) and another one for the intensity (Affymetrix) or intensityratio (cDNA chips). Because importing data is covered very thoroughly in the online manual accessible from *Help->Online documentation*, we do not go into details here.

### 1.1.2 Background subtraction

GeneSpring subtracts the background from the foreground or spot intensities automatically, if the background intensities are present in the datafile.

### 1.1.3 Calculation of expression change

For cDNA microarrays this is one normalization option, Per Spot: Divide by control channel, but the option is not used with lowess-normalization. There are two possible transformations, log-transformation (log of ratio) and fold change. The transformation options can be accessed from *Experiments -> Experiment interpretation*.

For the Affymetrix chips, a calculational control channel is created. The expression is then calculated as with cDNA microarrays. It is also possible to calculate the expression change using a certain control chip. This is one of the normalization options, Per Spot: Normalize to control samples.

**Table 1.1:** *Time series experiment*

	Time point	Replicate
chip 1	1 hour	1
chip 2	1 hour	2
chip 3	2 hours	1
chip 4	2 hours	2
chip 5	3 hours	1
chip 6	3 hours	2

Calculation of the expression change is coupled to normalization. Normalization only affects the control channel, which is adjusted so that the median intensity ratio of the chip is one when using Per chip: Normalize to percentile normalization. Therefore, the control channel is created for the Affymetrix chips, also. Note, that the original raw intensity values are always retained.

#### 1.1.4 Replicates

When importing the data from a chip where the same spot is present in multiple copies, GeneSpring automatically calculates an average of those replicates. This is assuming that the replicate spots have exactly the same Gene identifier in the data file. Replicate chips are also averaged, after defining the replicates in the *Experiment -> Experiment parameters* window and setting up the parameters in *Experiment -> Experiment interpretation*.

For example, if we have a time series experiment (Table 1.1) with three time points and two replicates per every time point, the parameters should be set up as follows. Two parameters, a time point and a replicate are created. In the time point, mark the time points suitably. The replicates are then set up using the other parameter. Inside one time point, the replicates are marked with a running number. Last, set up the value order for the time points. This tells GeneSpring, in which order the time points should be displayed on the screen.

If there is missing data for some genes either when importing the data or when setting up the replicate chips, GeneSpring only uses the existing data for the calculation of means.

#### 1.1.5 Checking linearity

Linearity is easily checked in GeneSpring using the M versus A plot. Go to scatter plot (*View -> Scatter plot*). Change the display options (*View -> Display options*) so that the horizontal axis is the average of raw and control (A), and the vertical axis is normalized (M). To help plot the non-linearity, from the Lines to Graph tab tick the Line of best fit box, which draws the linear regression line to the plot. Note, that the M versus A plot should be initially produced for unnormalized data.

### 1.1.6 Normality

Normality can be checked using the histogram. A histogram can be investigated in (*View -> Graph*) window by checking *All Samples* - interpretation mode, that is automatically produced for each experiment.

### 1.1.7 Filtering

A filtering tool can be invoked from the menu *Tools -> Filtering and statistical analysis*. A new window opens, where one genelists and one experiment should be selected. The actual filtering tool is accessed by right-clicking on the selected experiment and selecting *Add expression percentage restriction* from the opening list. Often the bad quality data is first filtered out. After that, the not-changing genes are removed from the dataset.

Using the scatter plot tool, define the intensity value, under which you can't trust your data anymore on either raw or signal channel. For cDNA chips, this is often around 200-1000 and for Affymetrix chips around 200. In the expression percentage filtering, use this signal value as a minimum cut-off. Also select that it applies to all the conditions. Create one such filter for both channels. Create a new genelists of the results (*Make list* button).

In the next filtering phase we will try to find the genes that are changing and we can trust in. Using the genelists created above, set up a new filter, where you select genes, which have expression values between 0.5 (minimum) and 2 (maximum). These genes are not showing any expression changes and are uninteresting to us. This filtering should also apply to the whole dataset (all conditions). After saving the new genelists, the filtering tool can be closed.

Go to the navigator bar and right-click on the good quality gene list. From the list, pick *Venn diagram -> Left (red)*. Similarly, add the not-changing and all genes genelists to the Venn diagram. Then select the *All genes* genelists from the navigator. From the Venn diagram identify the region that contains the genes included in the reliable genelists, but not in the not-changing genelists. Right-click on that area of the diagram, and make a list of these genes.

Now you have a list of genes, which are reliable and also changing. You're ready to proceed with the analysis, for example to clustering or classification analyses.