

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: Image analysis

Antti Lehmussola, Katja Kimppa and Pekka Tiikkainen

1.1 Image analysis using commercial software

1.1.1 ScanArrayExpress

ScanArrayExpress(SAE) software comes with the PE scanner, and is available in certain centers in Finland. SAE is scanning and image analysis software.

Version tested: 2.1

Environment tested: XP, 1024 Mb memory, 2 Ghz Pentium 4

Image analysis starts by opening images, one for each channel. If the images are opened simultaneously, the reference image assigning has to be checked (File -> Set control image) to ensure the calculations of ratios to be done correctly.

Quantitation of the image starts by selecting an existing quantitation protocol or using Easy Quant (define parameters for one analysis only). New quantitation protocols can be done through Configure -> Quantitation Protocols -> Add. When using an existing quantitation protocol, select the quantitation protocol and fine tune the locations of subarrays and images before starting the quantitation. When using the Easy Quant, user must define the grid either by inputting the needed values or by using a Gal-formatted file, and place the grid correctly on top of the images. The spot quantitation method and normalization method must be chosen also.

In SAE, four different spot quantification methods are implemented, fixed circle, adaptive circle, adaptive threshold and histogram methods.

- Adaptive circle: The program tries to find the edges of a spot and draws a circle around the spot. What inside the circle is the spot, and outside is the background.
- Fixed circle: The program draws a specified size circle (specified in the template) around the spot, and what's inside the circle is the spot, and outside is the background.
- Adaptive threshold: This method uses statistical algorithm to define the spot.
- Histogram: In this method the histogram of spots is calculated and used to define the spot and the background.

Status	Flag
Found	0
Not Found	1
Absent (call from GAL file)	2
Present	3
Bad (only by user interference)	4

In SAE only local background is calculated.

In SAE two different normalization methods are implemented, linear normalization method and LOWESS normalization. These are applied to the whole dataset.

SAE uses 5 different classes to classify the spots, these classes correspond in the result file Flag column values as follows:

SAE gives out 49 columns for 2 channel data, including spot locations, gene annotations, raw information of the spots in individual channels, statistical information about the spots, calculated ratios and normalized data. SAE calculates both mean and median values for spots, backgrounds and ratios.

SAE has got some tools for preliminary checkup of the data, including few scatterplot tools f.ex. MA-plots or intensity-intensity plots drawn from raw or normalized data, and distribution plots.

After quantification, SAE shows the data as a spreadsheet and the grids located over the images. In this step user has got the possibility to modify the automated results, moving spots or changing the status of the spot. The selected spot is shown selected in each of the different views to the data.

Pros

- Easy to use
- Versatile possibility to adjust grids and spots
- Interactivity after the automated analysis
- Possibility to change f.ex spot locations after analysis
- Multiple options for spot analysis method

Cons

- Only local background is calculated
- If the hybridization is bad, the automated spot finding may drift from correct location
- For array size of 30 000 spots, the analysis of one array takes 30-45 minutes
- The response to the commands in adjustment of grid locations is slow

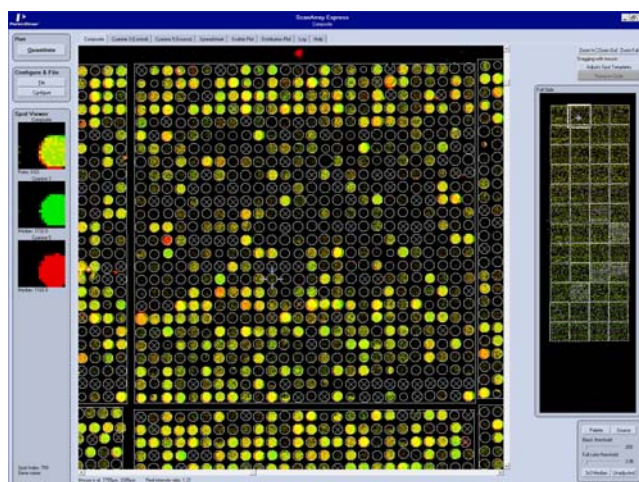


Figure 1.1: Screenshot of SAE after spot finding. Note that the data selection is transparent through all of the windows (composite, cyanine3, cyanine5...) simultaneously. Three images of the array are also available at all times, the main analysis image (in the middle), full image of the array (on the right) and a close up for a spot (on the left).

1.1.2 Agilent Image analysis

Agilent Image Analysis (AIA) software is distributed with Agilent microarray scanners, and is also available for 30 day test use through the Agilent web-pages. Image Analysis software is available in centers which have got Agilent microarray scanner.

Version tested: A.7.5.1

Environment tested: XP, 1024 Mb memory, 2 Ghz Pentium 4

AIA imports only Agilent format TIFF-files, in which one file consists of stacked images for both channels. The analysis flow is the following: crop the image, place the grid on top of the array, defining the parameters and saving results.

The grid information can be imported to the AIA from Agilent's grid file, GAL-formatted file, Agilent design file, tab text file or manually. The software shows the inputted values for the grid and after that the grid may be fitted on top of the image manually or automatically. After this it is possible to fine tune the locations for whole grid, sub grids and individual spots. The grid has to be saved and after that the feature extractor can be called, in which we have to define which modules need to be ran and define the parameters for the modules, like the spot analyzing method, background subtraction method, normalization and quality measures. After this we can run the feature extractor. When it's ready, all the results will be saved automatically and it shows the visual results on top of the image. All outliers (uneven intensity or background, uneven shape or population outliers) are shown with differently colored circles on top of the spots. Dark blue circles are non-outliers.

In AIA there is two different ways to analyze spots, cookie cutter and whole

spot methods. Cookie cutter is a circle based method and the whole spot method bases on finding the edges of the spot.

Six different background calculation methods are implemented in AIA including no background subtraction, local background, average of all background areas, average of negative control features (needs information of the negative controls), minimum signal on feature and minimum signal on feature of background.

On normalization, there is linear and lowess methods, and combination of these two. In here it is also possible to select which spots are used for normalization.

On AIA you can look at intensity distributions across the slide (line of pixels) or defined smaller area, like unique spot.

AIA gives out 5 different result files: GEML, MAGE, JPEG, Text and visual results. The text file includes the results in tab-delimited format, including raw data (intensities, backgrounds, means, medians etc) and calculated values (all intermediate values from data processing steps). Depending slightly on the options selected, there are 90-100 columns of data. Note, that AIA logarithmic values are 10-based logarithms.

Pros

- Easy to use
- Versatile possibility to adjust grids and spots before spot finding
- Multiple options for spot analysis method
- Multiple options for background calculations

Cons

- Takes in only Agilent TIFF-files
- For array size of 30 000 spots, the analysis of one array takes 20-30 minutes
- No possibility to change spot calls or locations after analysis

1.2 Freeware image analysis software

A wide variety of software for microarray image quantitation is available on the Web. For an academic user, paying thousands of euros for a license of commercial software might not be possible. For the average mortal - with limited budget funding - using a free-ware program is often the only option.

This last part of the chapter represents three such programs. In addition to being free to use, they all had to fulfill two additional criteria: independence of any other software (e.g. Matlab) and ability to run in Microsoft Windows. All programs are meant for quantitation of spotted arrays. For a comprehensive list of available quantitation software, see http://ihome.cuhk.edu.hk/~b400559/arraysoft_image.html.

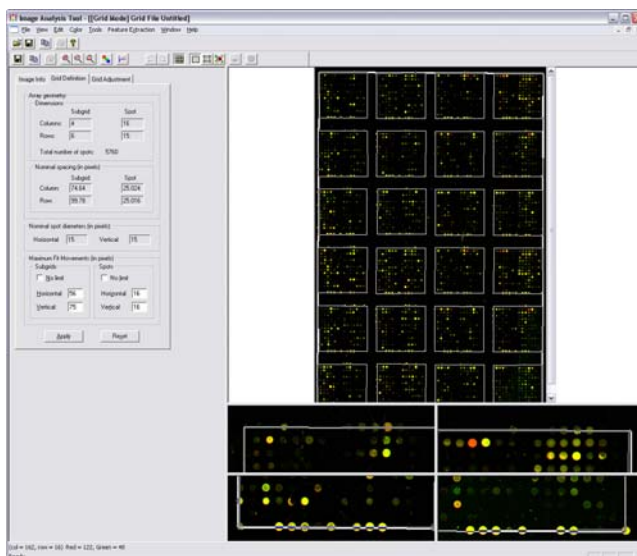


Figure 1.2: AIA in action: grid adjustments. Circles over the spots include quality information in color coding.

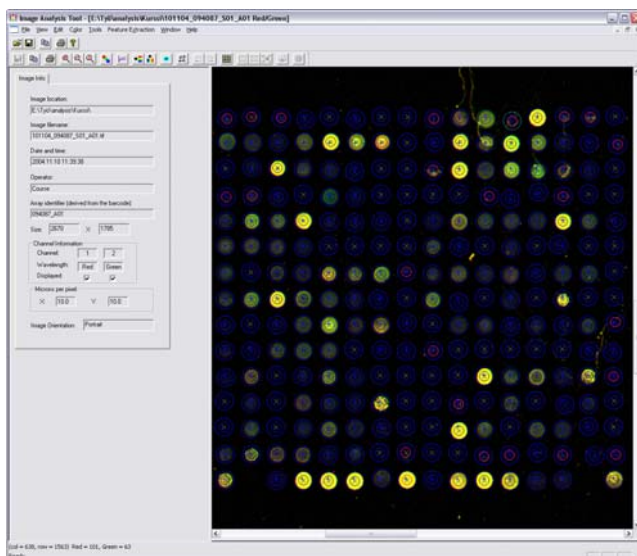


Figure 1.3: AIA in action: spots after analysis. Circles over the spots include quality information in color coding.

1.2.1 TIGR Spotfinder

Version tested: 2.2.3

Available at: <http://www.tigr.org/software/tm4/spotfinder.html>

The process to quantitate images with Spotfinder is quite straight-forward: load the images, create subgrids and overlay them on the subarrays in the image,

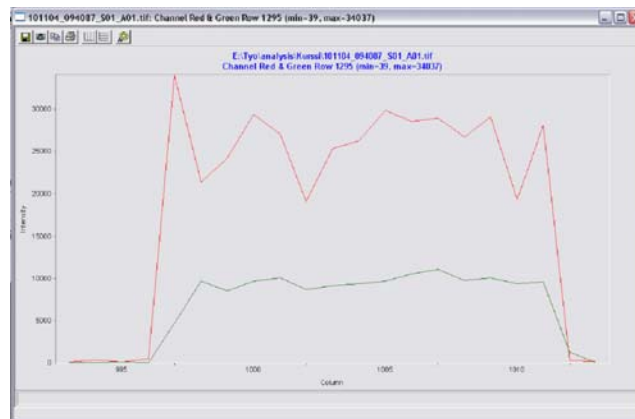


Figure 1.4: AIA view of spot intensity distribution.

process the grids and export the data. Overlaying the subgrids, or putting them on the right place, takes some 20-30 minutes for a large 20,000 spot image. If the subarrays aren't perfectly in right angle with the image, the grids can be tilted to fit them better. A pattern identification algorithm then automatically locates the borders of the spots and calculates both foreground and background intensities. Each spot is given a quality control value based on its form and intensity. This parameter can be later used for example in determining present spots. Also included is a possibility for some quick light-weight data analysis.

The learning curve for using the program is short thanks to the intuitive user interface. Earlier versions of the program used to have bugs that caused crashes demanding the user to save his work frequently. The version tested gives a few warnings every now and then but is stable.

Pros

- Quick to learn, fast to use
- Subgrids and cells can be twisted and resized manually for optimal results
- Automated spot identification
- Intuitive user interface enditemize

Cons

- Exporting directly to Excel works only for individual subgrids
- Only local background can be calculated

1.2.2 ScanAlyze 2

Version tested: 2.50

Available at: <http://rana.lbl.gov/EisenSoftware.htm>

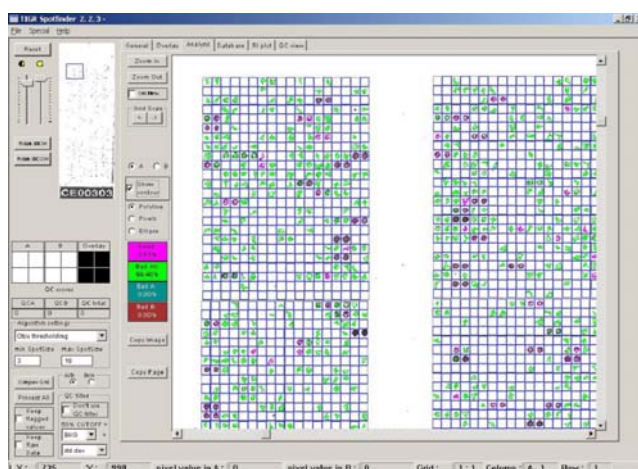


Figure 1.5: AIA view of spot intensity distribution.

The first thing the user notices of ScanAlyze is its weird user interface. Each function from image loading to grid creation has its own window. A special Window control box is used for opening other windows. With a normal screen resolution (1024 x 768) all contents of most windows cannot be viewed at once requiring scrolling the windows up and down. All this makes using ScanAlyze a painful experience. It is nice that software designers try different approaches in user interface design but in this case originality comes with a price.

After loading images for both channels, the user needs to create the spotting grids. Strangely, only grid structures of 1, 4, 16 and 32 subgrids are available. For quantitating arrays with a different array format this is a major obstacle requiring unnecessary tricks to get the job done. Moving the grids is difficult as drag and drop with the mouse works only for individual grids. Also zooming in and out in the image is very slow. Black image background makes it very hard to determine subarray boundaries if they are closely packed and/or most spots have low intensity. Position for several spots needs to be individually assigned which means that a lot of time is required for quantitating a large array.

The program assumes that all spots are round which is not always true. Everything inside the circle is considered foreground which causes biased results if the assumption of roundness is not fulfilled.

Pros

- Easy to use auto-uninstall utility included

Cons

- Poor user interface
- Low intensity spots not shown properly when zoomed out

- Number of subgrids not freely adjustable
- Spots are assumed to be round
- Zooming and moving around the image is slow
- Quantitation takes time due to the amount of manual work needed

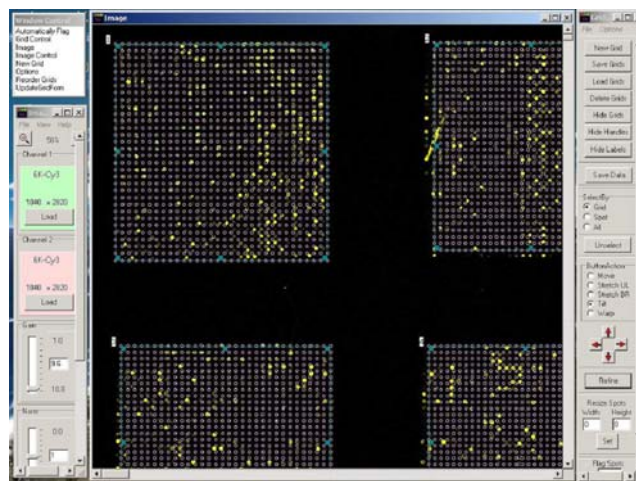


Figure 1.6: AIA view of spot intensity distribution.

1.2.3 GridGrinder

Version tested: 1.3

Available at: <http://gridgrinder.sourceforge.net/>

Makers of GridGrinder claim that it can automate the process of array spotting and quantitation. All the user needs to do is to tell which images to process and give some information of the subarray architecture. After doing this, you should be able to just lay back and watch the computer crush some numbers. Not only saving time and labor, this should also result in more reproducible outcome since the program does its job more objectively than human. Unfortunately, the reality isn't that care-free.

The user interface is very clear and therefore easy to learn. The images to be processed are loaded in a list, output folders set, the arrays' architecture defined and the processing started. Depending on the size and number of arrays, processing takes anything from few to several minutes on a normal desktop computer. In addition to the numerical output data file, the program also produces a copy of the original image files showing the spots it recognized. They can be used for refining parameters if the program failed to identify all spots correctly. The actual output file contains plethora of information about the spots and their foreground and background stats. That should satisfy even the most demanding data analyst.

When tested with a medium-density array with 6,144 spots, the program was able to find the location of all spots correctly after a few tries. Processing one image took one to two minutes which is reasonable. Performance was much worse when tested with an image of a large array with some 20,000 spots. The task was especially hard since most spots had no or very low intensity values. This time the processing took closer to 10 minutes and the program failed to locate the spots.

In conclusion, the program works fine with smaller arrays which have most spots present. The performance gets worse when less spots are present and noise increases. After optimized for a specific array design, the program can speed up research to some extent. It is up to the researcher if he or she finds that improvement worth it. After all, most projects today include the use of 10–20 arrays. An experienced user can quantitate those easily in a day with semi-automatic software like TIGR Spotfinder.

Pros

- Plain and intuitive user interface
- Automated processing suits well for some arrays
- Can quantitate several images in a batch
- Output data includes a diverse set of measurements

Cons

- Performance degrades if noisy and dense images are used
- The time saved with automated processing is marginal for small microarray projects

Suggested reading

1. Gonzalez, R.C., Woods, R.E. (2002) Digital Image Processing, Prentice-Hall, New Jersey.
2. Russ, J.C. (1999) The Image Processing Handbook, CRC Press, Boca Raton.
3. Zhang, W., Shmulevich, I., Astola, J. (2004) Microarray Quality Control, John Wiley & Sons, New Jersey.

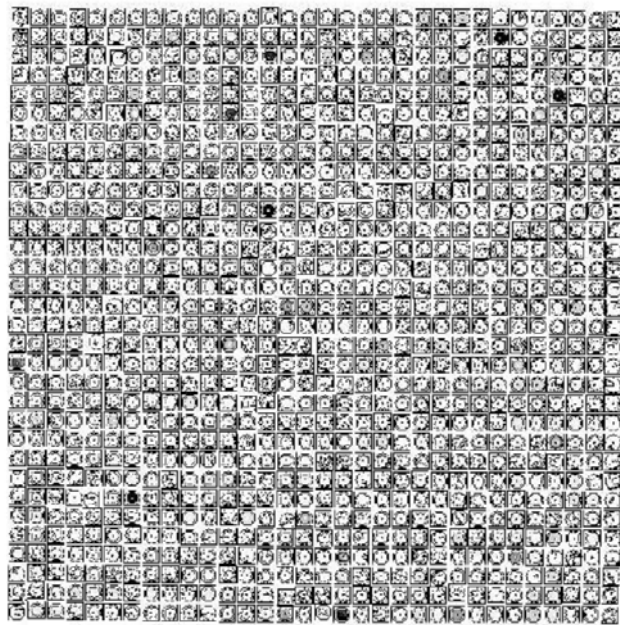


Figure 1.7: *AIA view of spot intensity distribution.*