Automatic gene expression estimation from microarray images

Daniel O. Dantas

Adviser: Junior Barrera

IME-USP

Summary

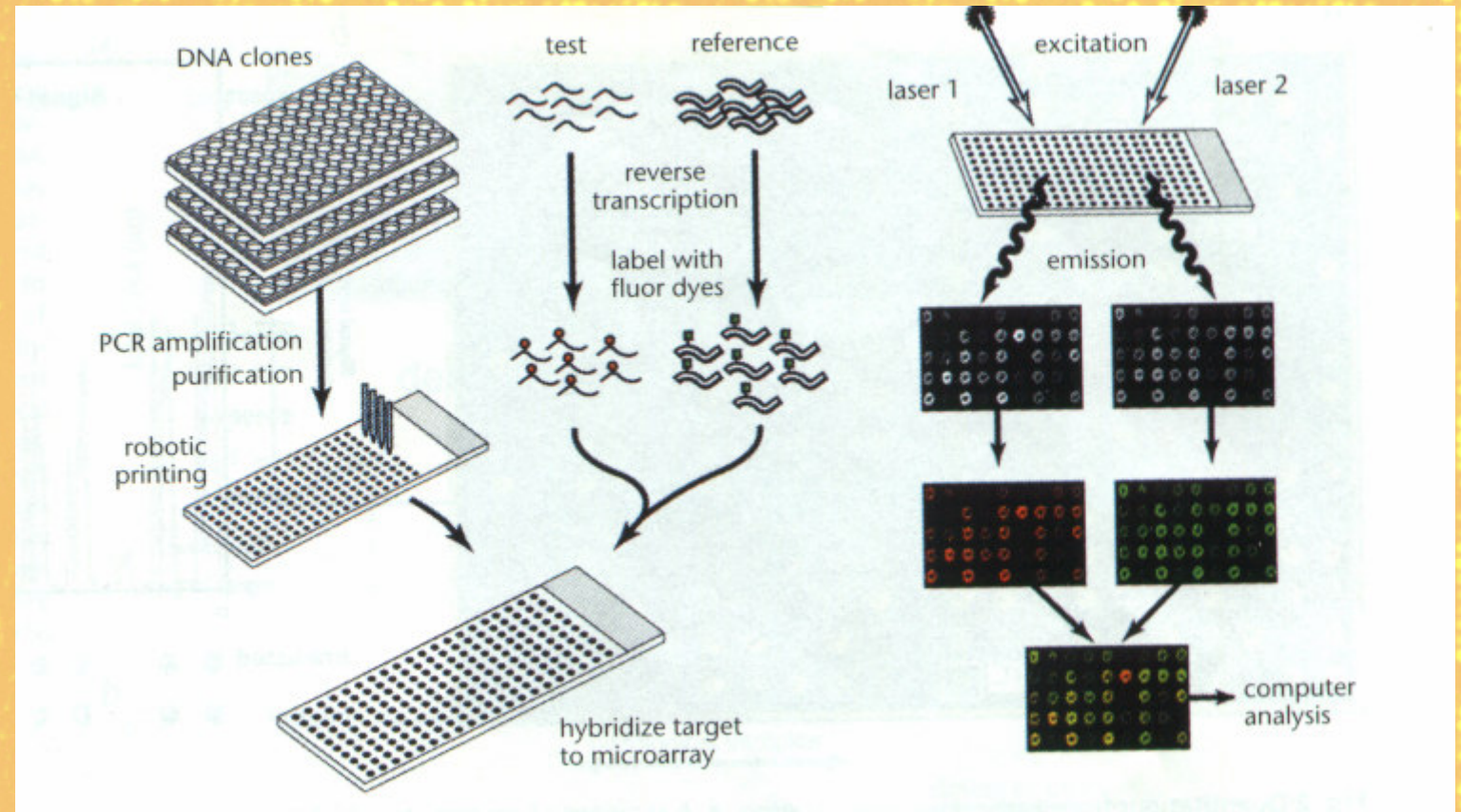
- Introduction
- Problem definition
- Solution strategy
  - Image segmentation
  - Signal estimation
  - Validation
- Conclusion

Introduction

- Microarray is a hybridization based technology used to measure the relative abundance of mRNA from two samples
  (cancer and normal tissue, bacteries under normal and stressing conditions)

- Hybridization = matching of pairs of nucleic acid

Data acquisition



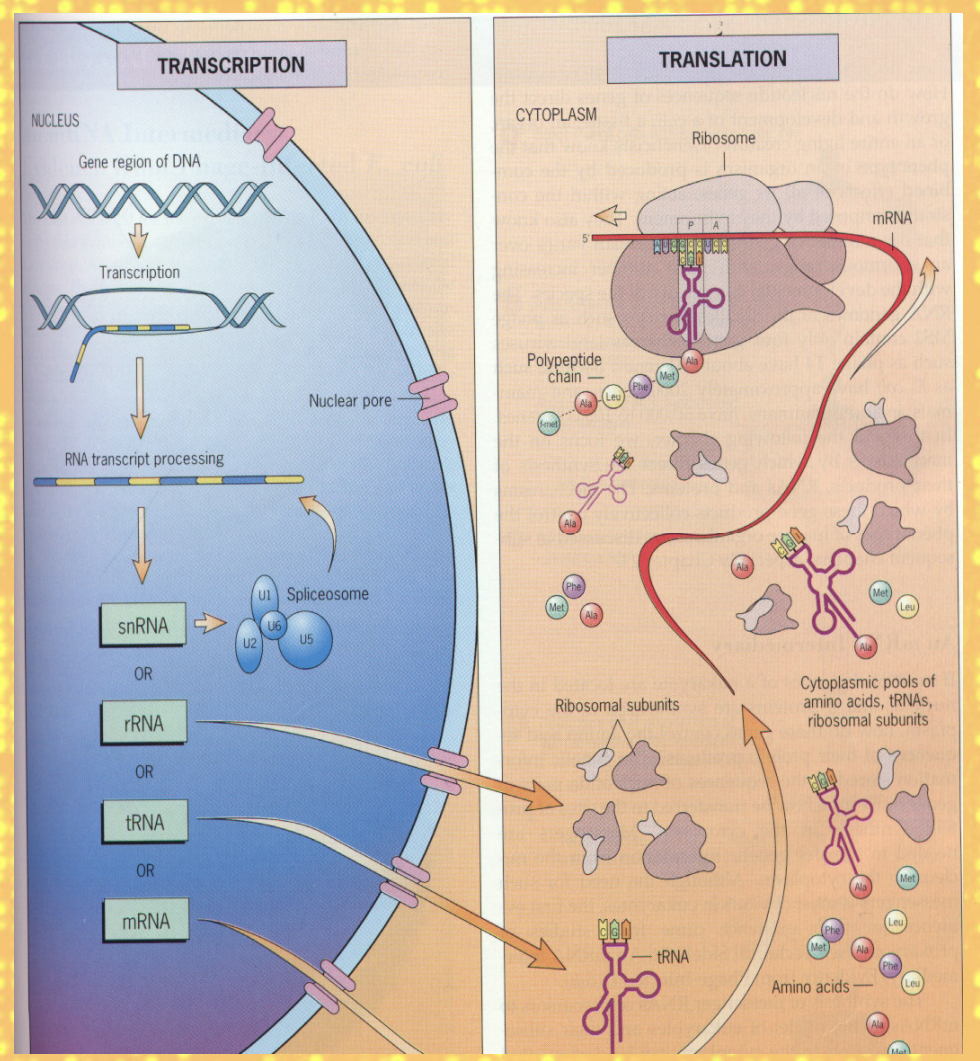
What is it for?

- Used to compare gene expression under different conditions.

- "Gene expression is the entire process that takes the information contained in genes on DNA and turns that information into proteins." (edtech.clas.pdx.edu)

Knowledge evolution in genetics

- Gene expression

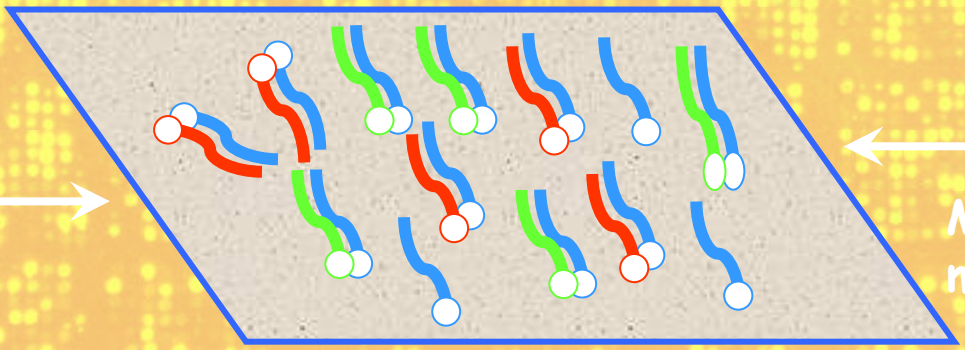


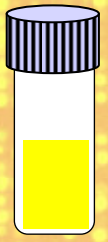
How does it work?

- Fix in a glass slide samples of cDNA.
- Extract mRNA from the two kinds of cells you want to analyze.
- Label copies of the mRNA from each sample with different fluorescent dyes.
- Pour the two soups onto the glass slide and leave it there for some hours.

Cinética de hibridização

cDNAs

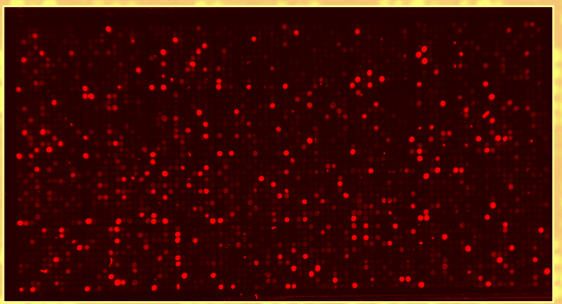




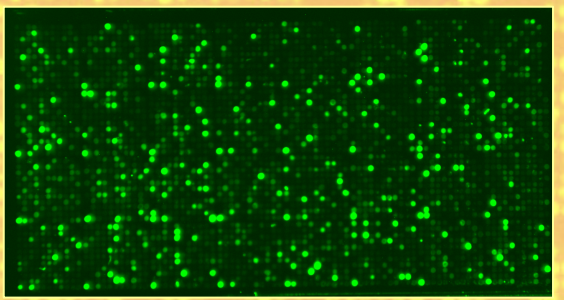
Misturas de cDNAs marcados *CY3* e *CY5*

Captura das Imagens

Cy5



Cy3



How does it work?

- If the mRNA finds a matching cDNA, they will hybridize. The more mRNA in a sample, the more the respective color will lit.
- The scanner measures the light emitted by the fluorchrome when excited by a light at an appropriate wavelength.

A scanned image of a microarray slide

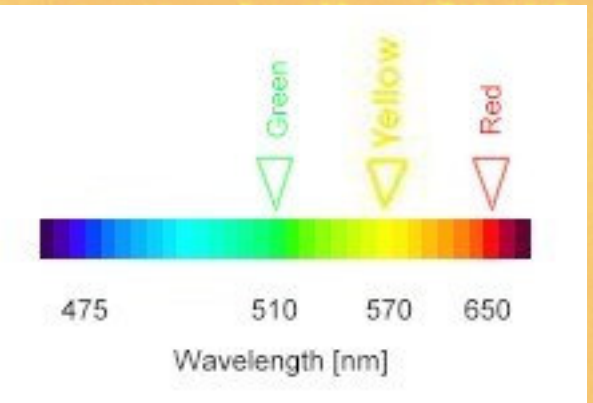
A microarray slide

- Is a small glass slide with about 1”x3”
- The resolution of a typical microarray image is about 10μm (1000 pixels/cm).
- Each pixel of one channel has 16 bits = 2 bytes (ranges from 0 to 65535)

2 bytes x 2 channels x 2000 x 4000 = 32MB

A microarray slide

- The red channel represents the cy5 (wavelength = 635nm)
- The green channel represents the cy3 (wavelength = 532nm)

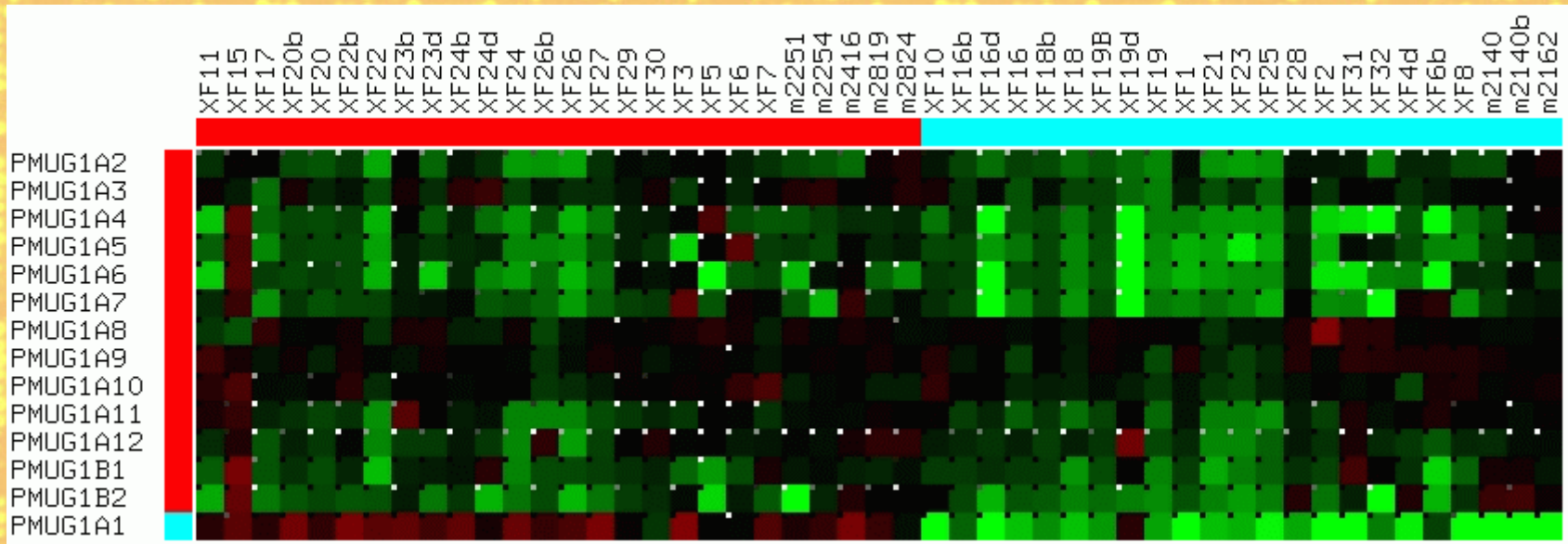


Problem definition

- Create a table with the estimated gene expression of each gene spoted in the slide *automatically and reliably*.

Aplication

- We can use the expression data to compare the behavior of many genes and classify them using clustering techniques, for example.



Available solutions

- Scanalyze: usually doesn't find misaligned spots.
- SpotFinder(TIGR): subarrays must be placed manually.
- Arrayvision: very good on locating misaligned spots; many options.
- UCSF Spot: does everything automatically if the image is perfect.
- Quantarray, F-scan, Dapple, Genepix, Imagene etc.

- All of them require user interaction to some level.

Our aim...

- Is to reduce the user interaction, doing the job automatically and measuring correctly the relative mRNA concentrations.
- This will make the process cheaper and faster.
- User interaction makes the segmentation subjective. Eliminating that, the results may be more reproducible.

Solution strategy

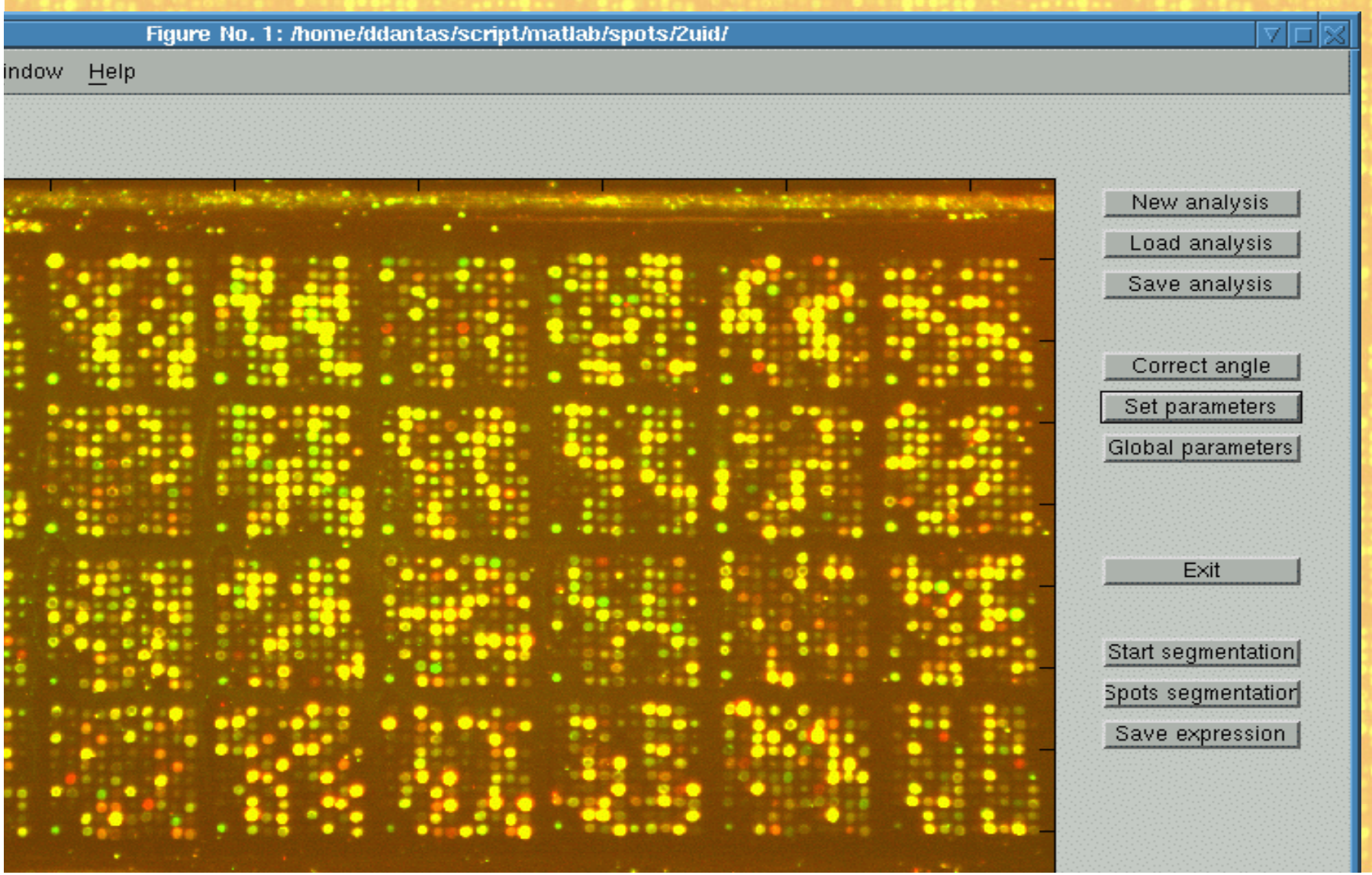
Manual steps

- Tilt correction (optional)
- Microarray geometry parameter setting

Automatic steps

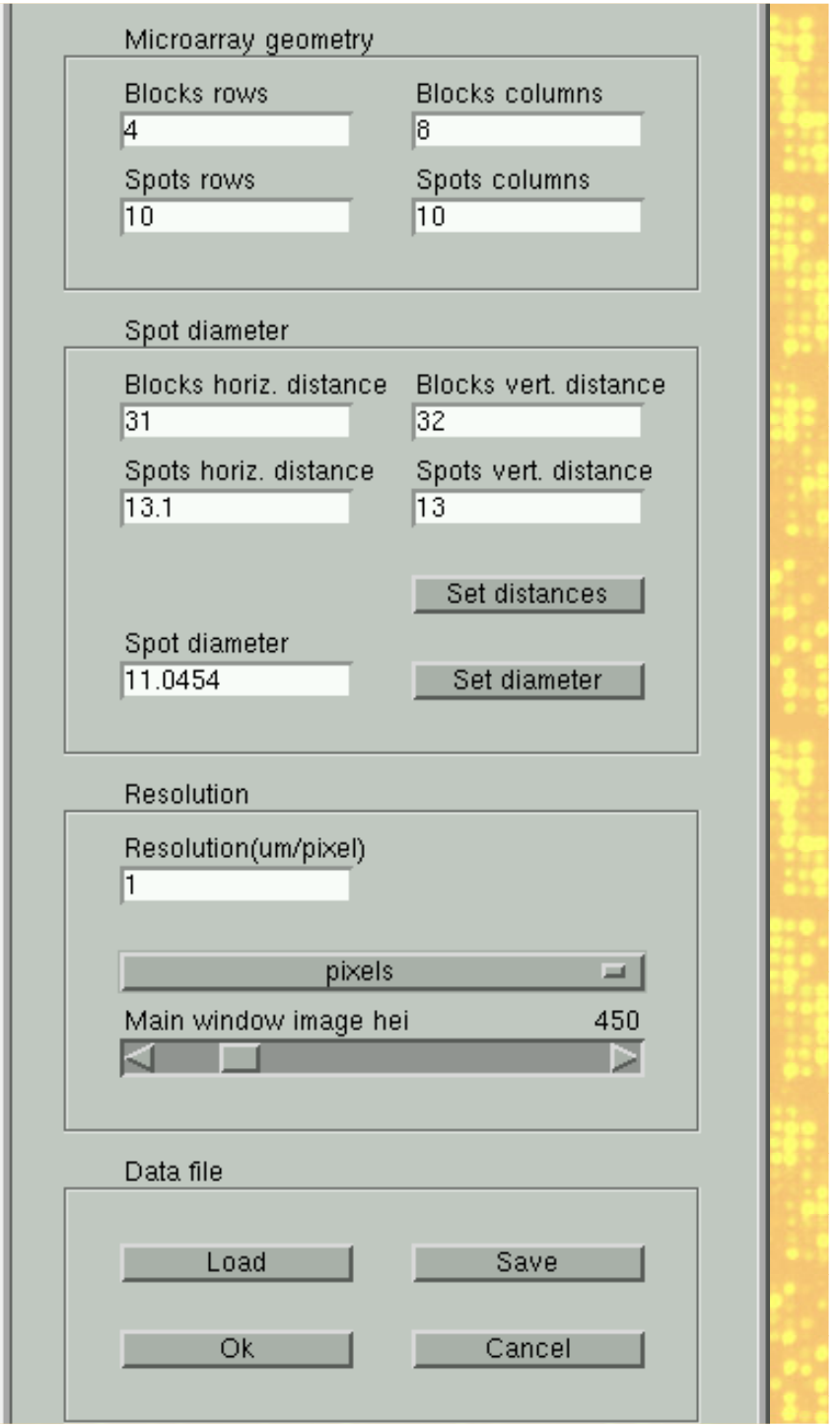
- Subarray gridding using image profiles
- Spots gridding using image profiles
- Spots detection
- Gene expression generation

Our software

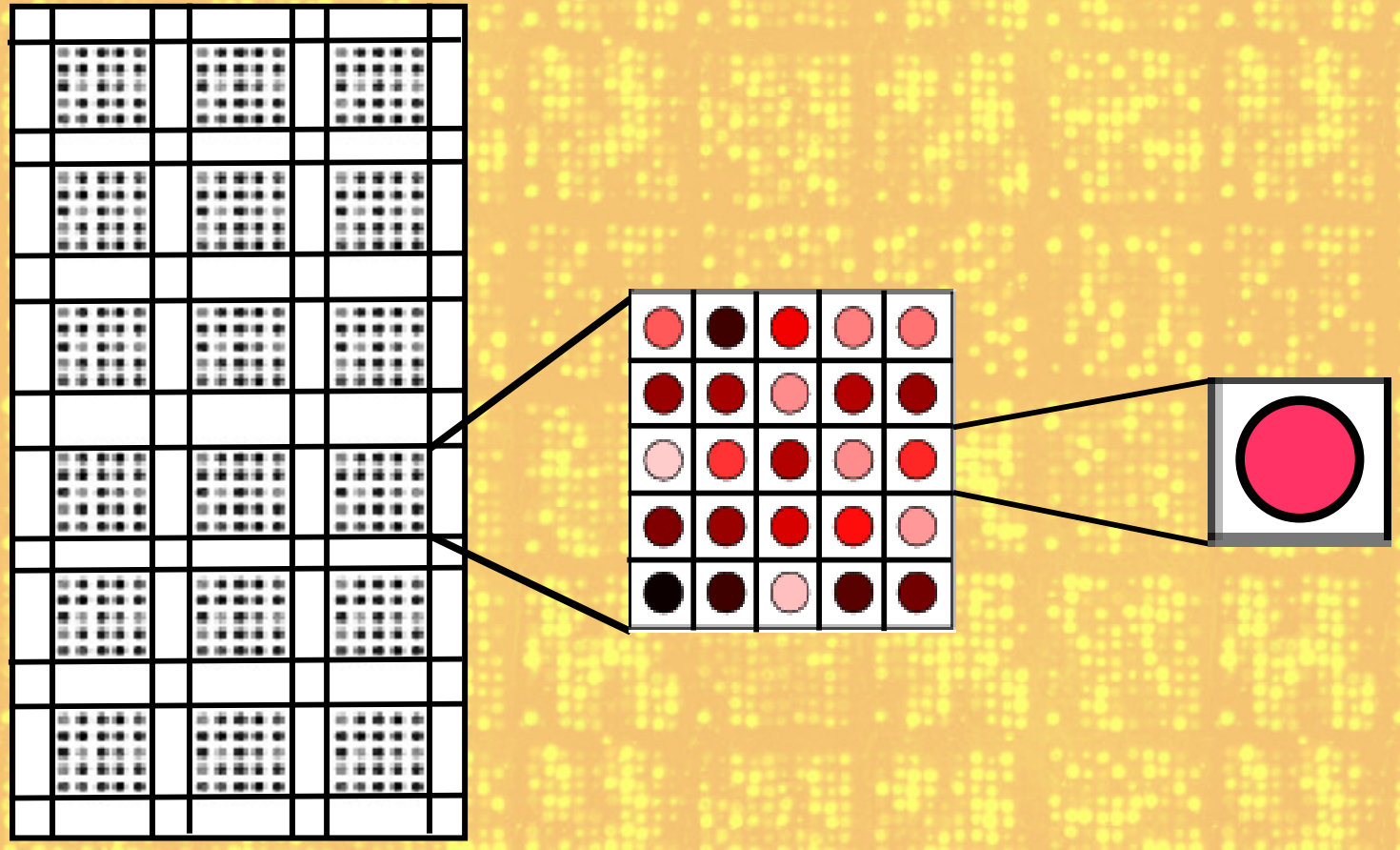


Parameter setting

- In this window the user sets parameters for a whole family of arrays
- He can save in a file for reusing them



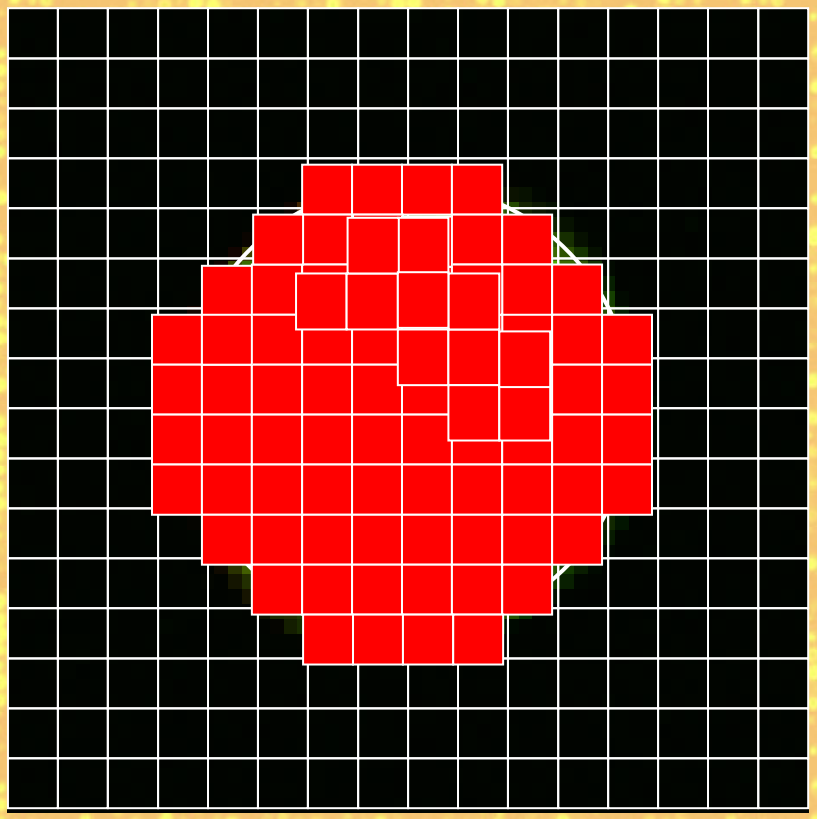
Microarray image segmentation process



Hirata R, Barrera J, Hashimoto R, Dantas D, Esteves G. *In press*, 2002.

Microarray image segmentation process

Delimited the spot, we must choose which pixels will be used in the signal estimation



We can use every pixel

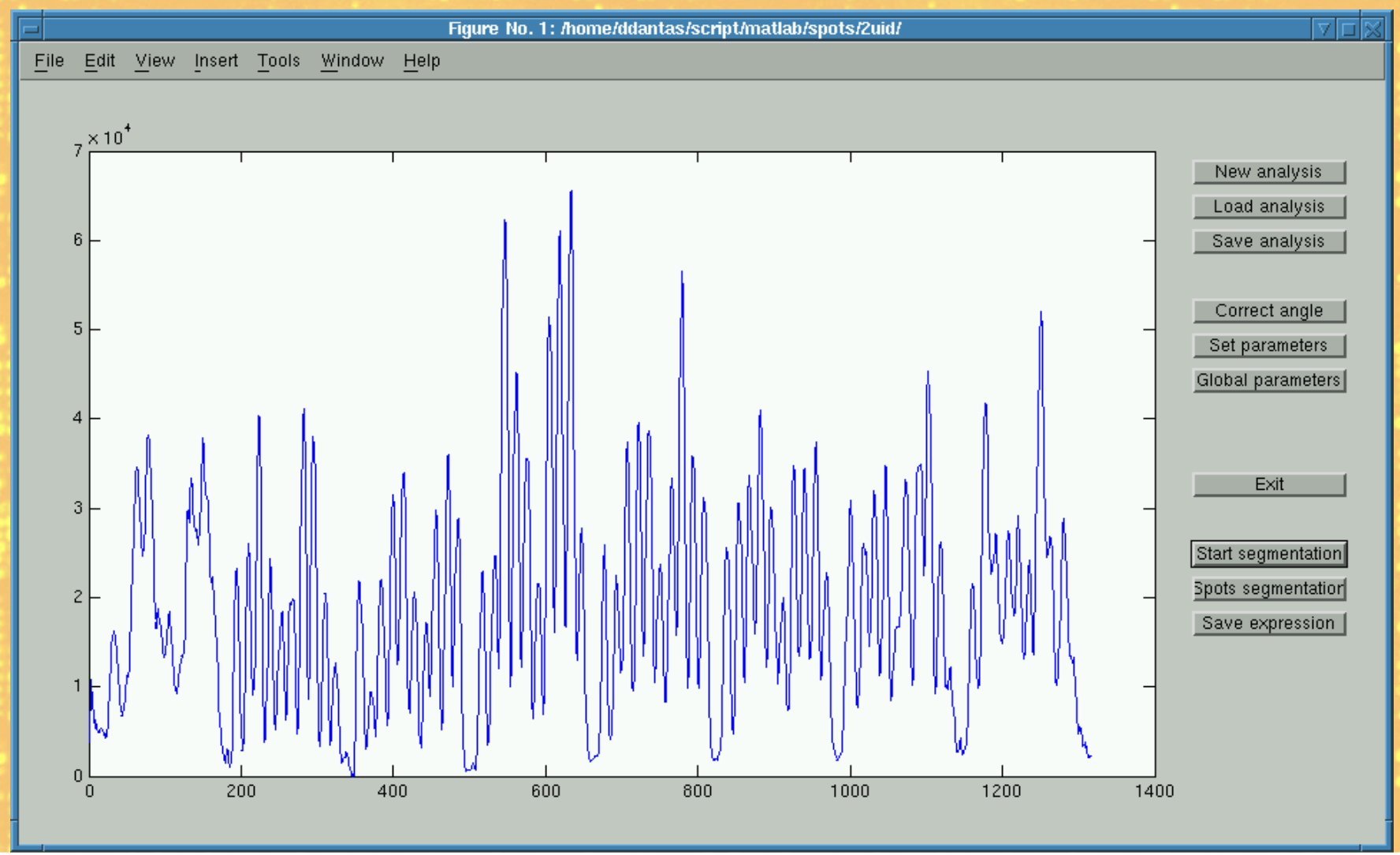
Or we can select some of them based on the histogram information

Example
15% of intensity of foreground

The same is done in the background

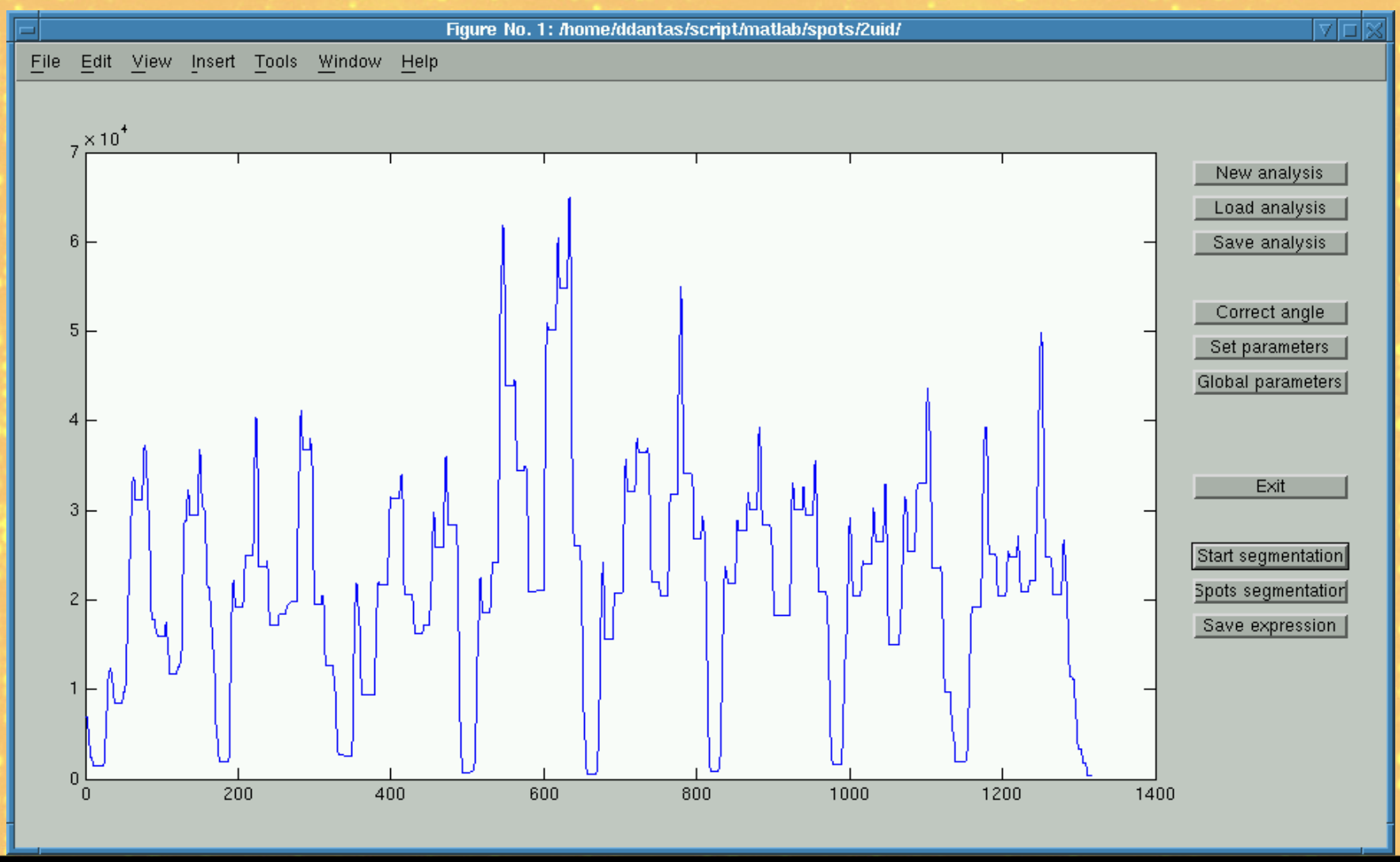
A vertical image profile…

is the sum of the spots values of each image line



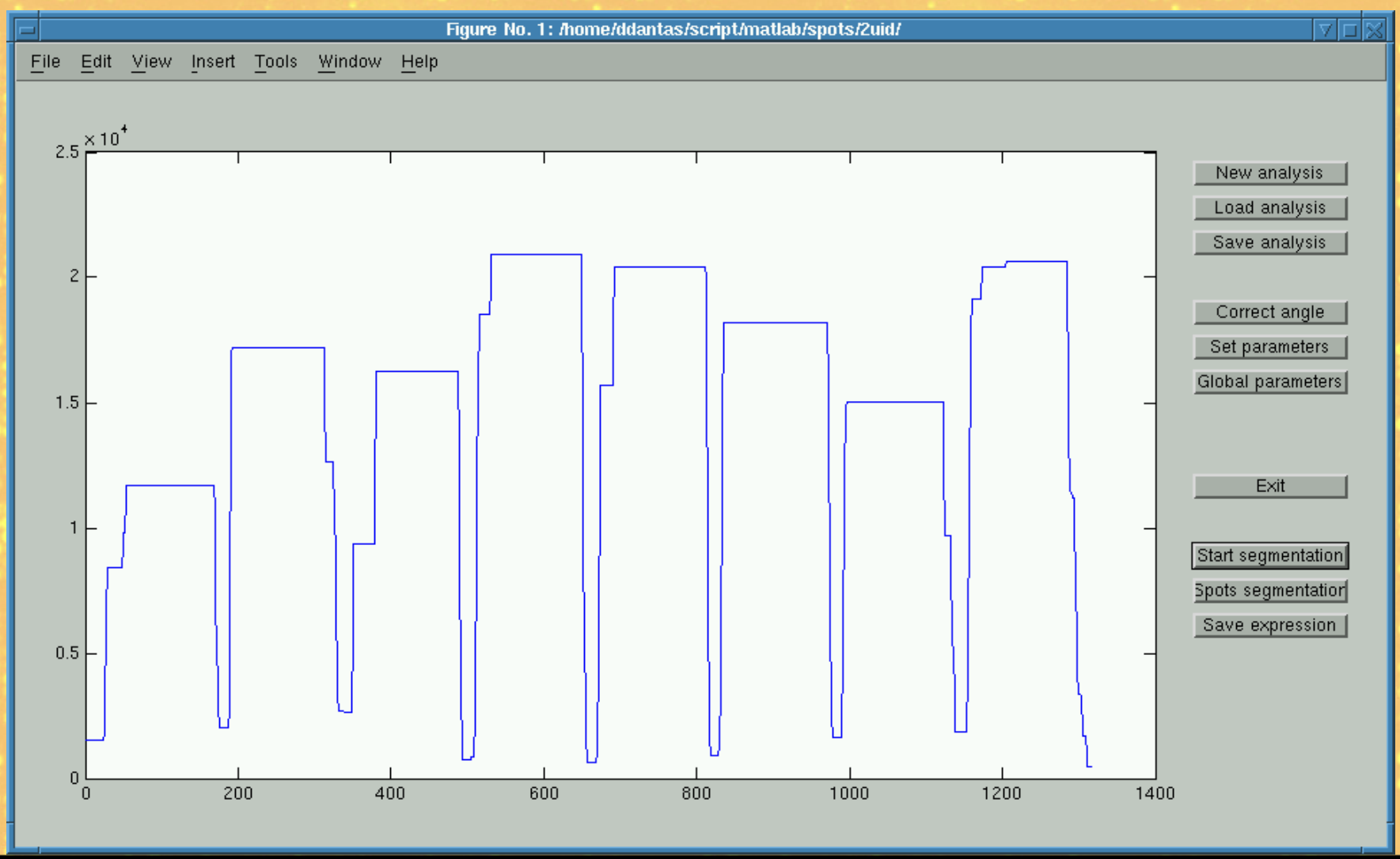
The subarray gridding…

Is done by filtering the horizontal and vertical profiles



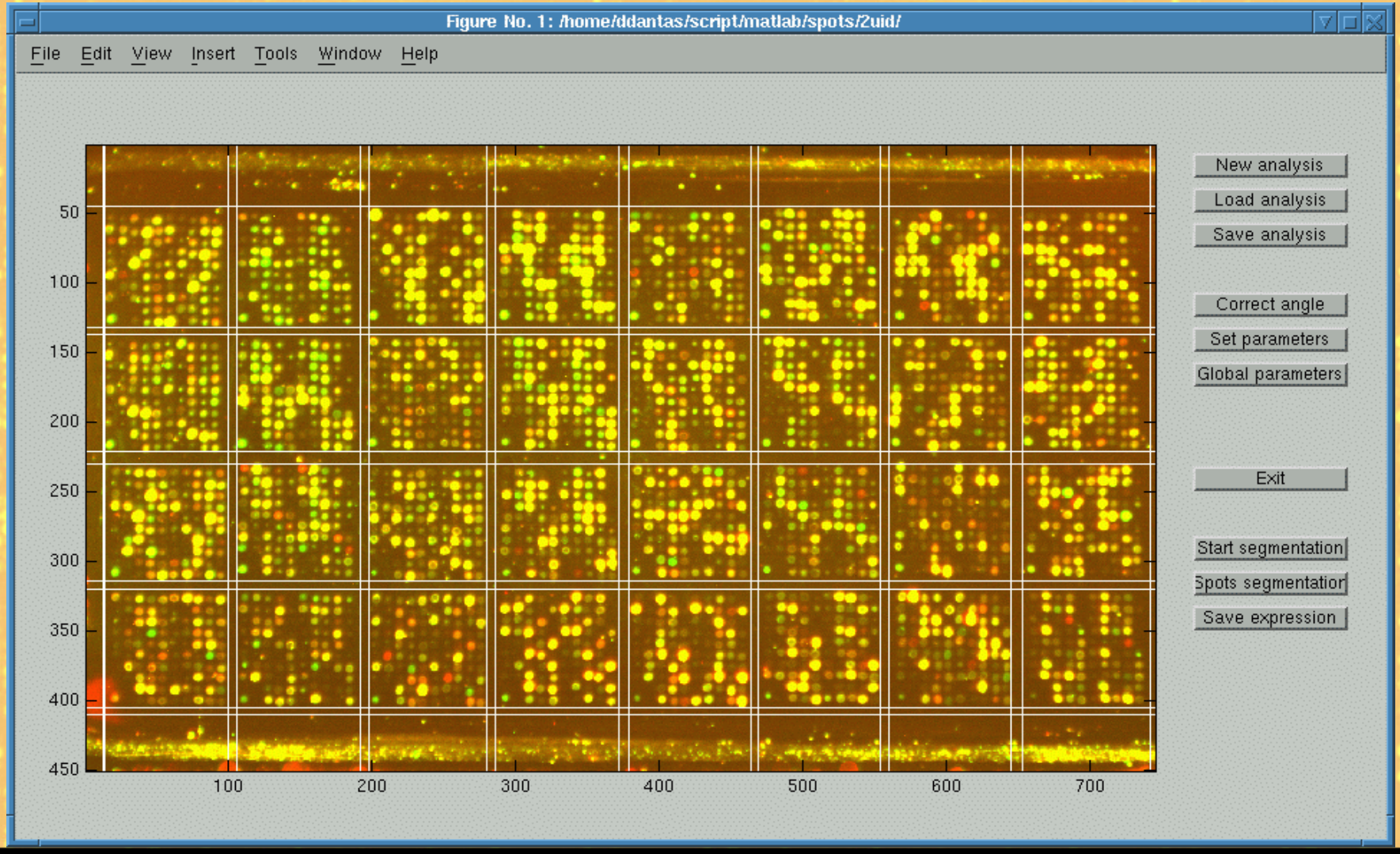
And finally…

taking the local minima of the filtered profile



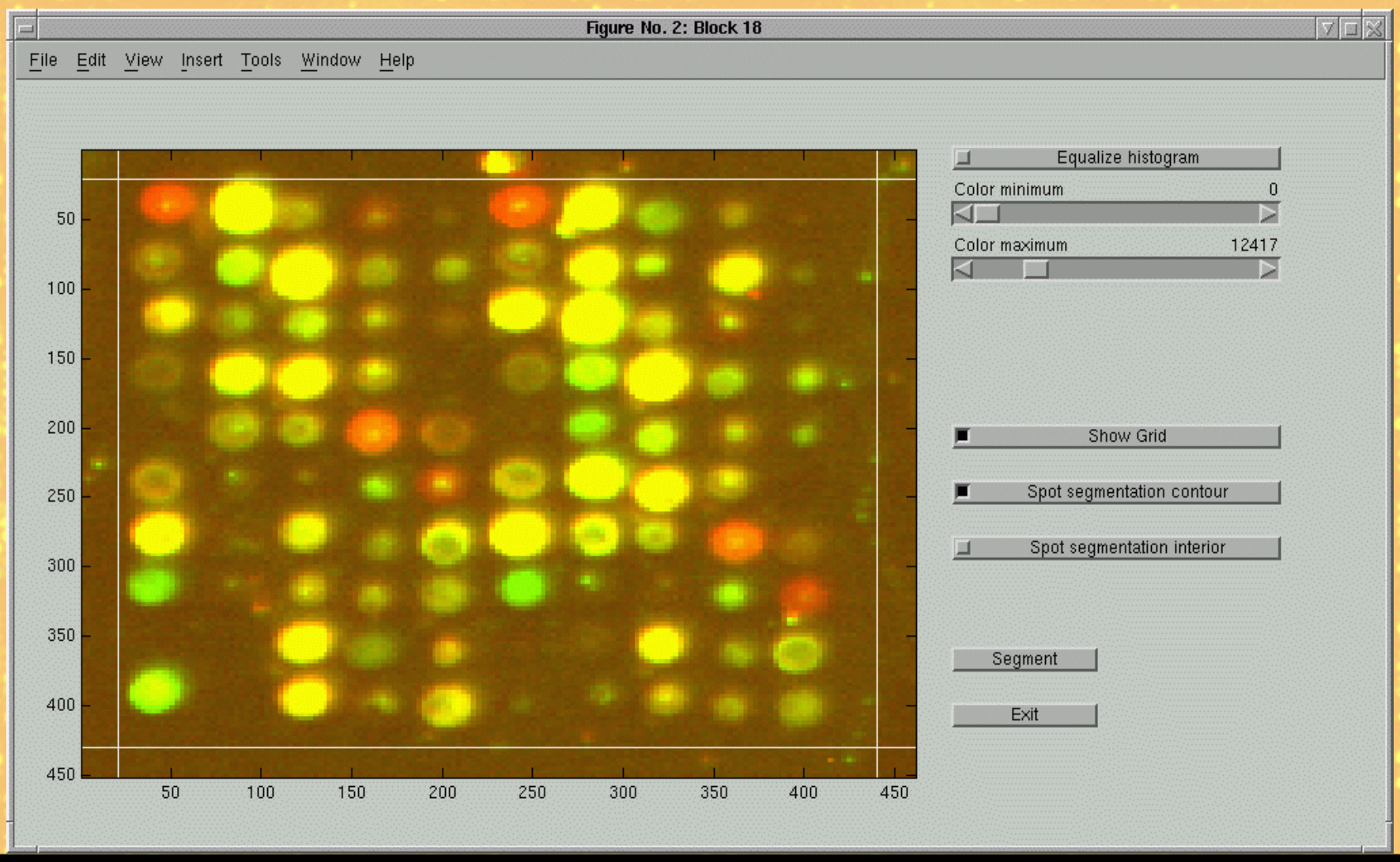
the same is done with…

the horizontal profile. Here the result



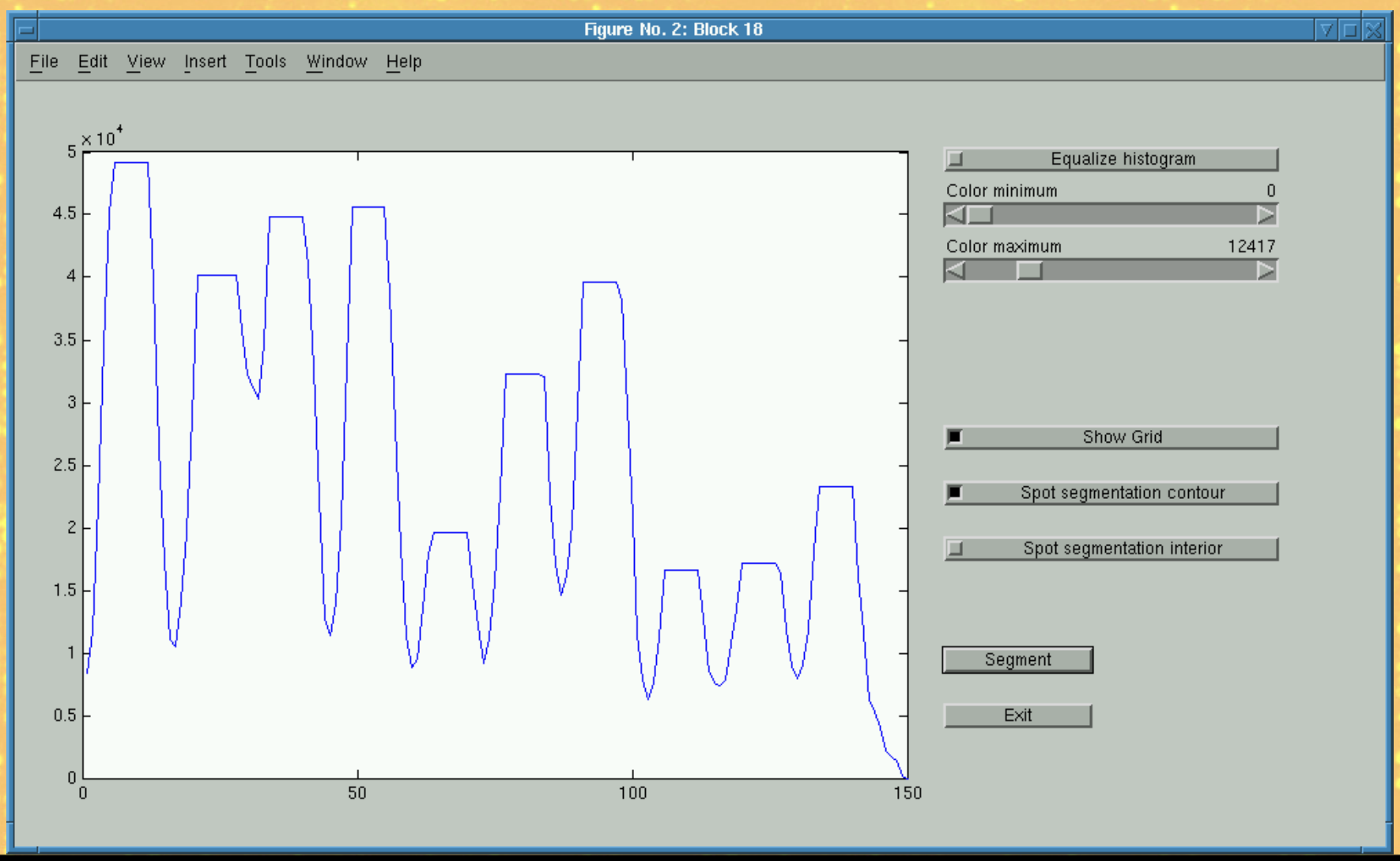
Spots gridding…

is done separately for each subarray



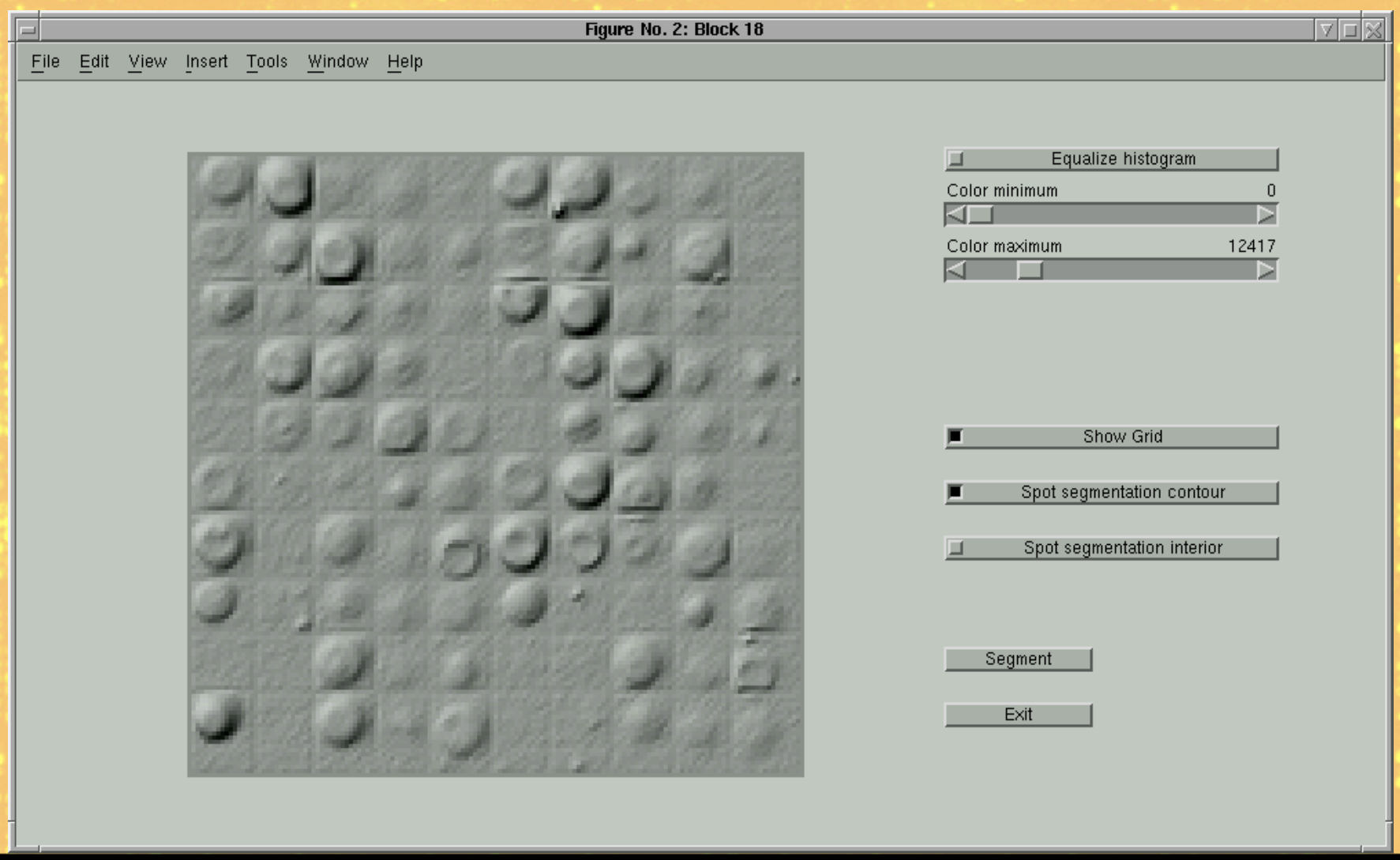
The profile filtering is simpler…

having just one step, and also uses local minima



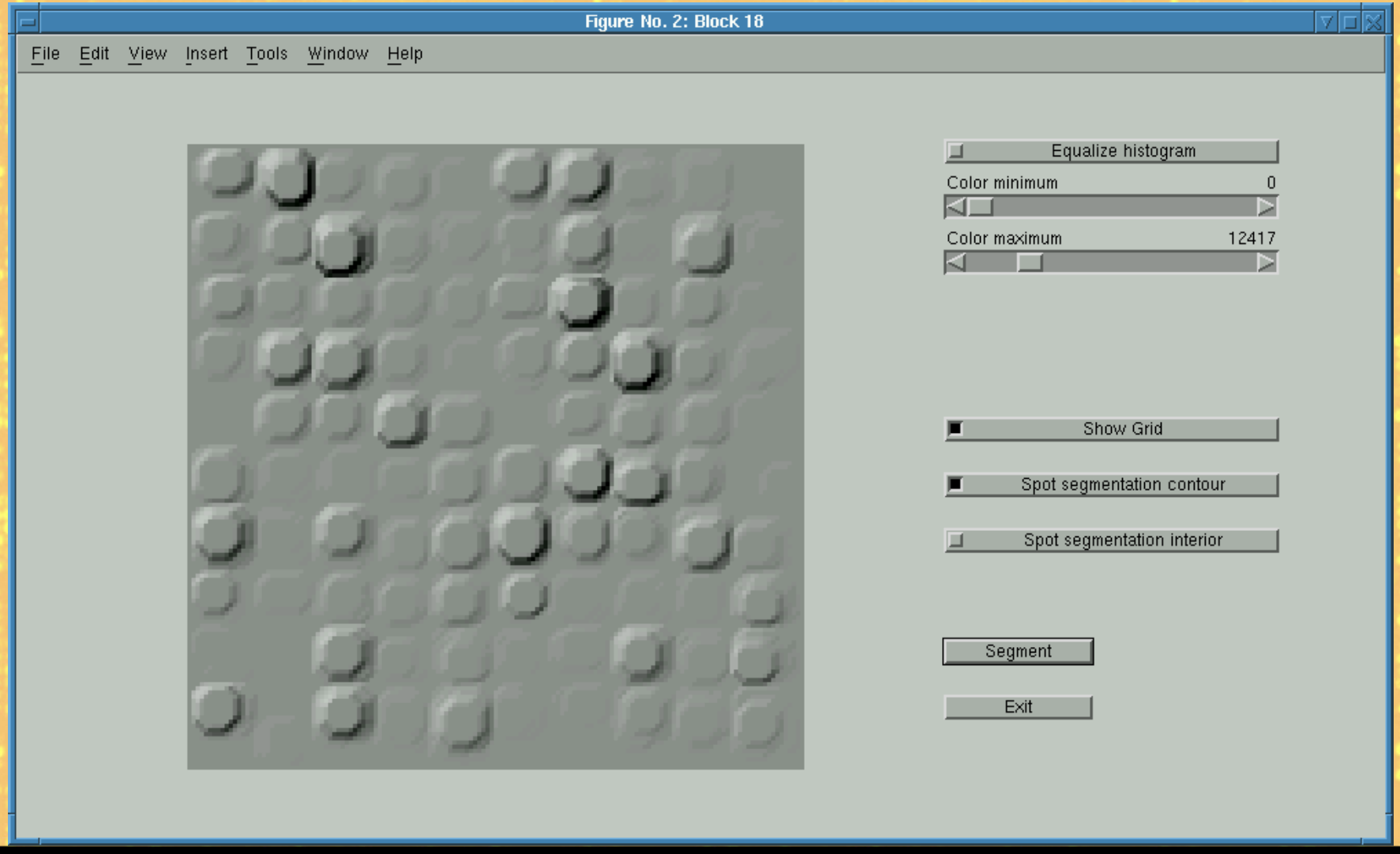
The spots detection step…

is basically the application of the Watershed operator



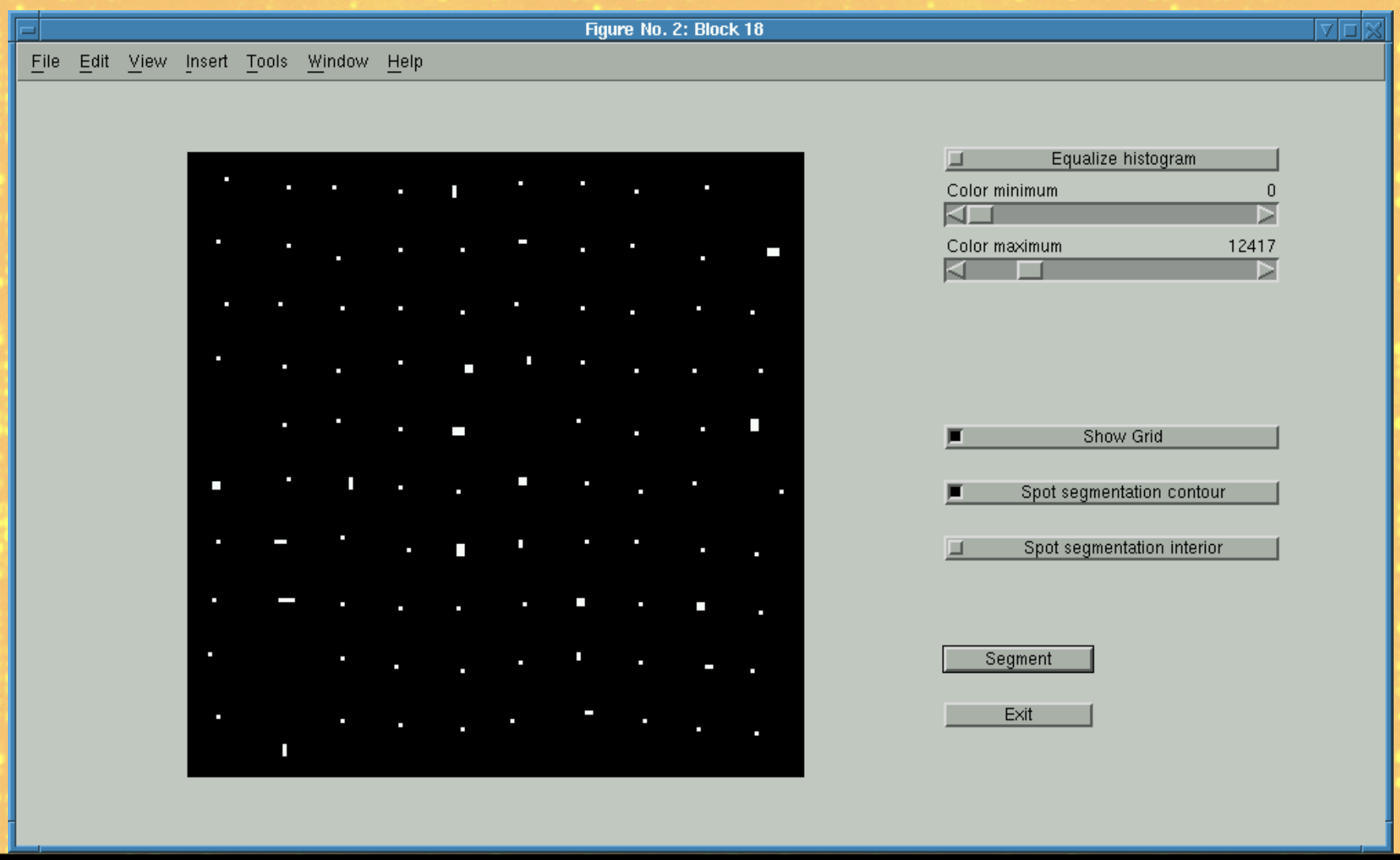
To avoid oversegmentation…

the image must be filtered



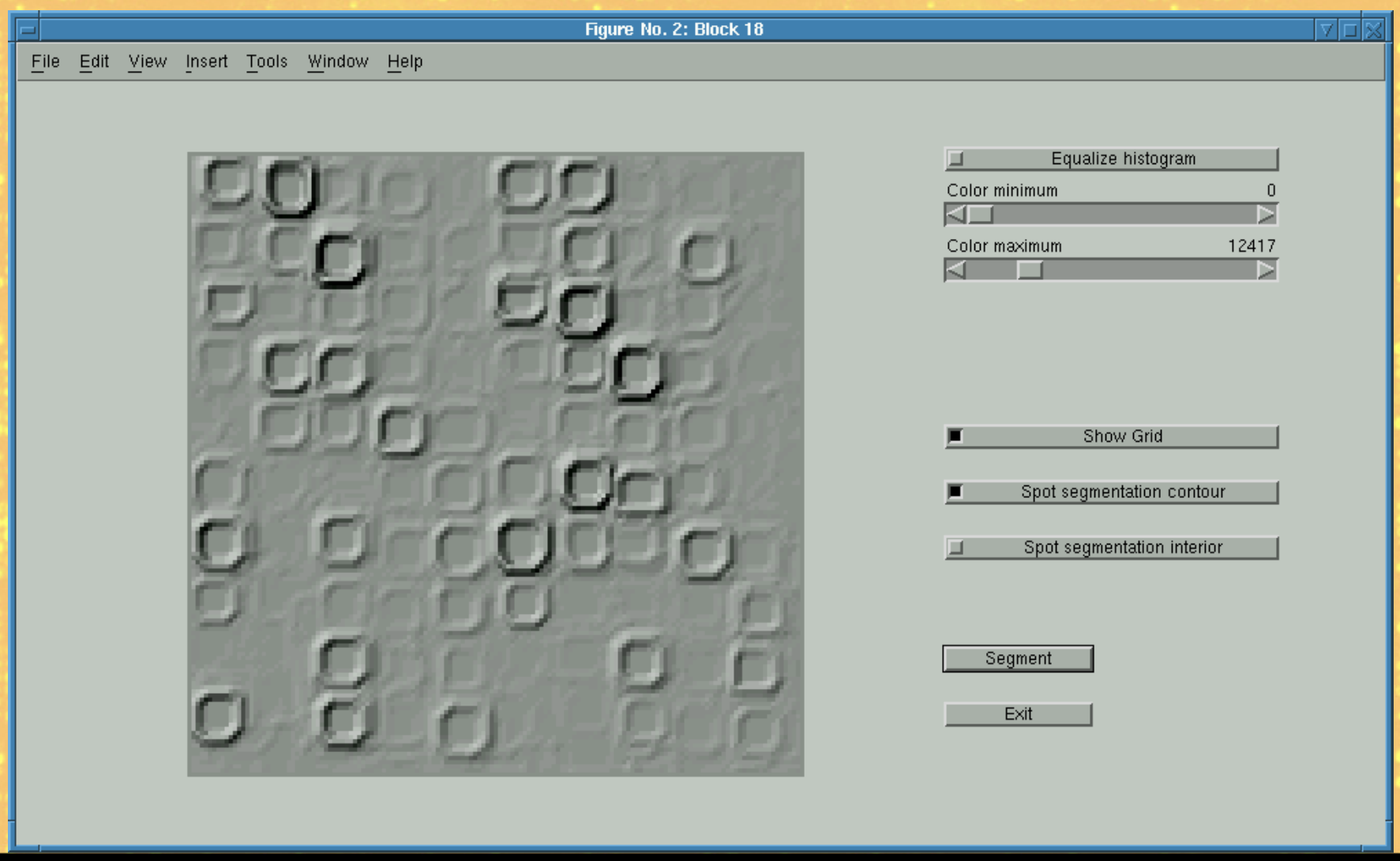
The filtered image also gives…

markers that will be used in Watershed



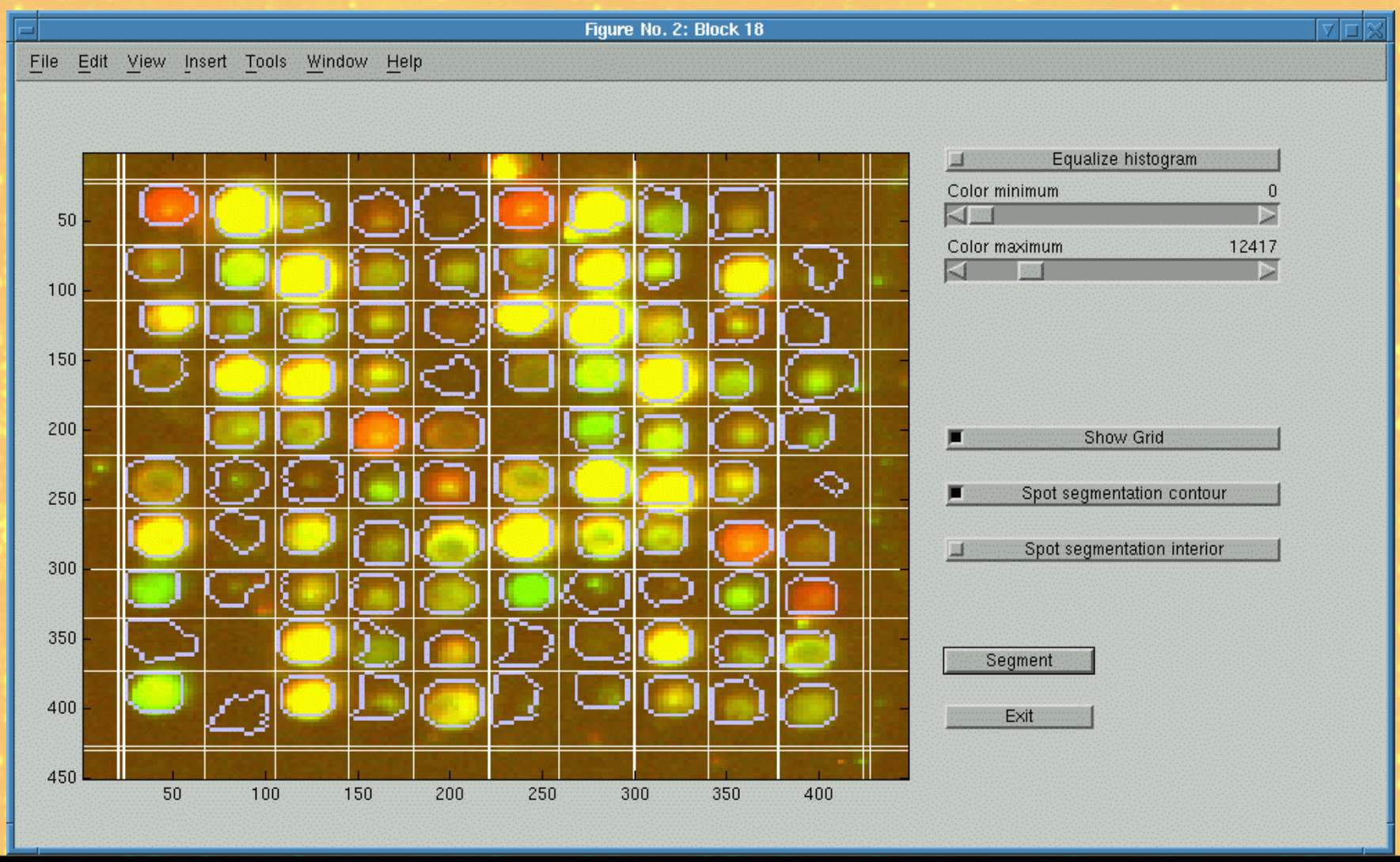
We give as input to the Watershed…

the markers, grid and the filtered image gradient

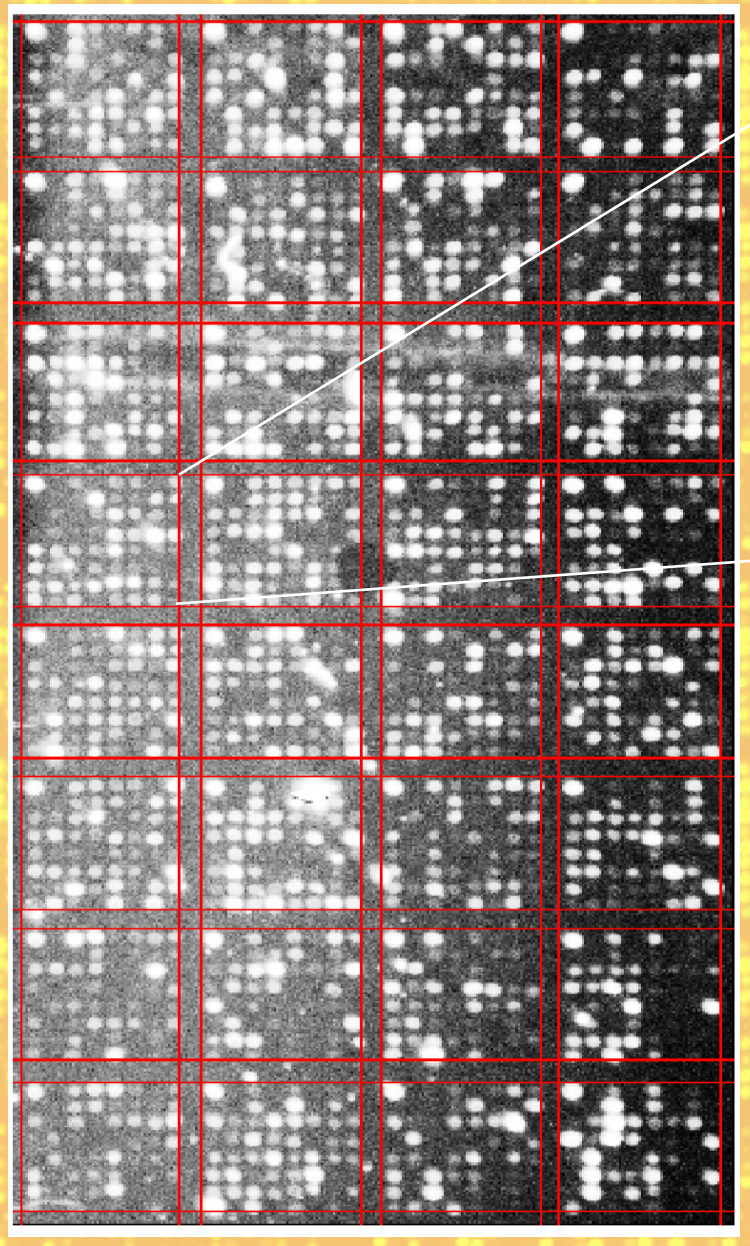


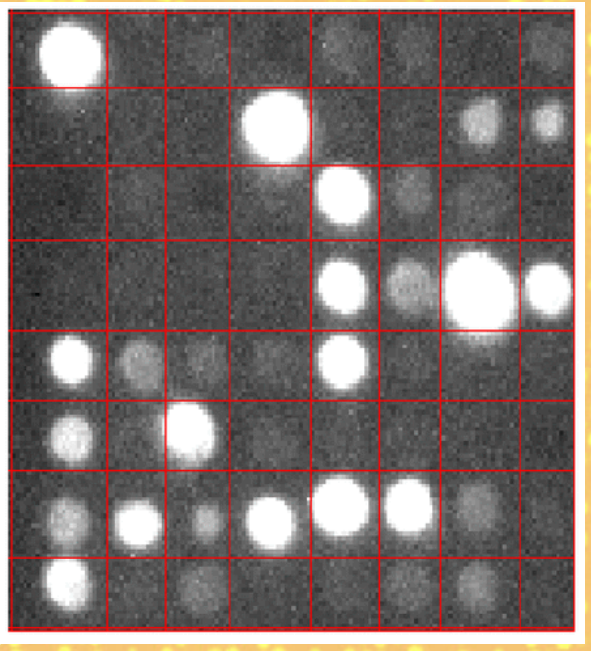
Here the resulting…

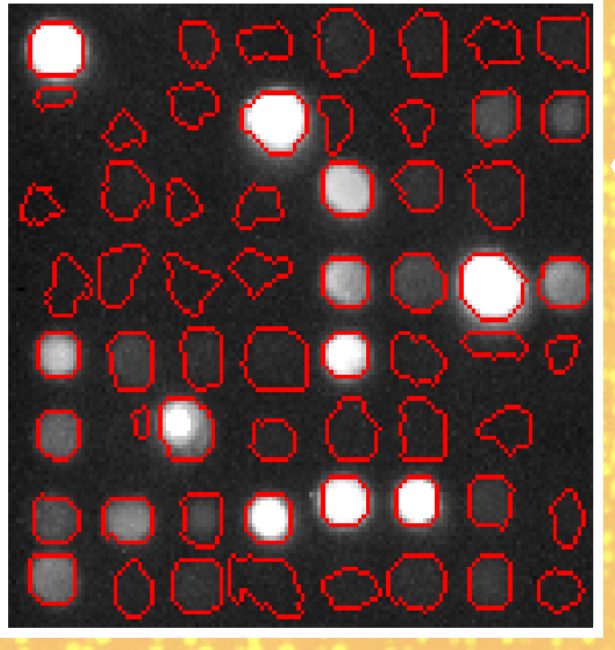
grid in white and spots cortours in light blue



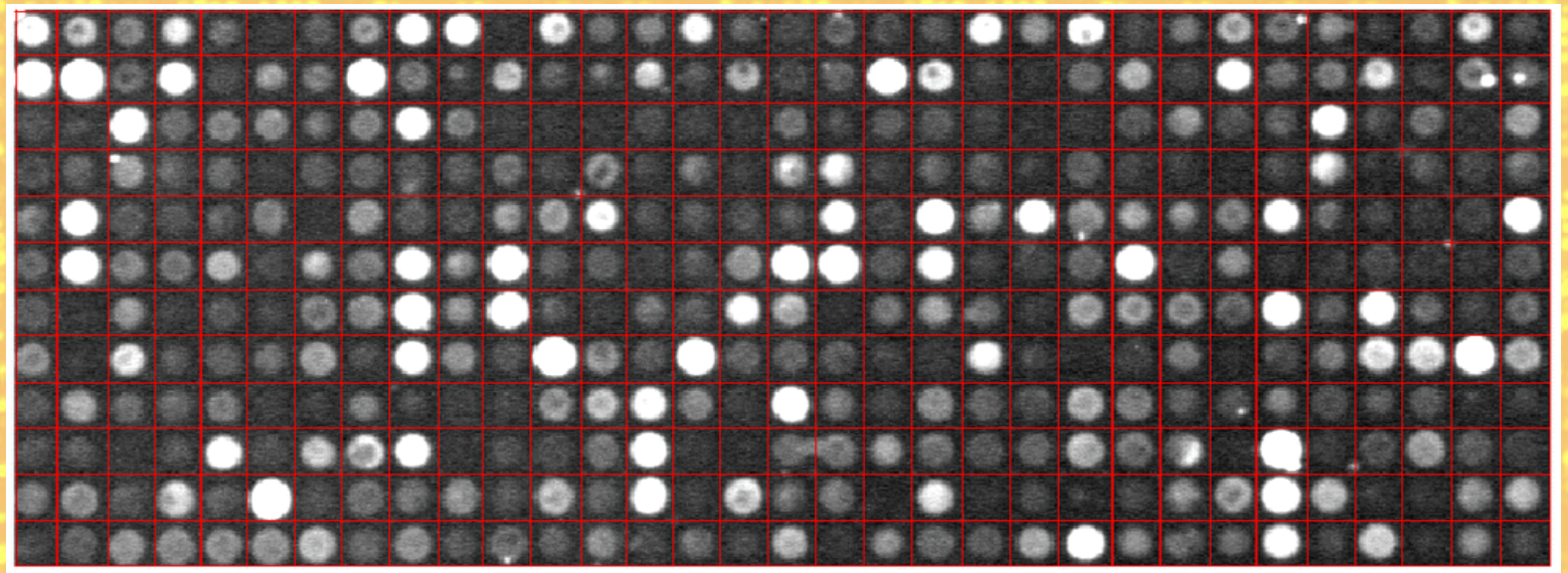
Segmentation example

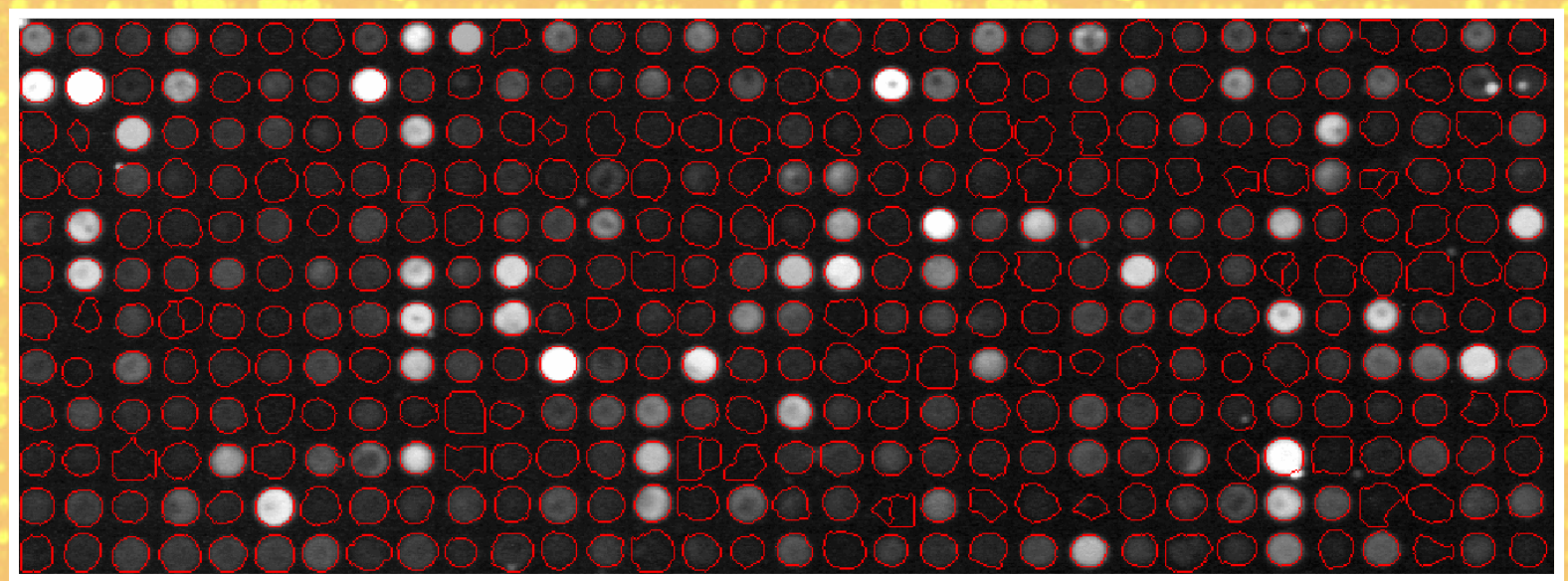






Segmentation example

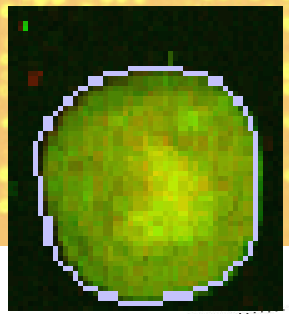


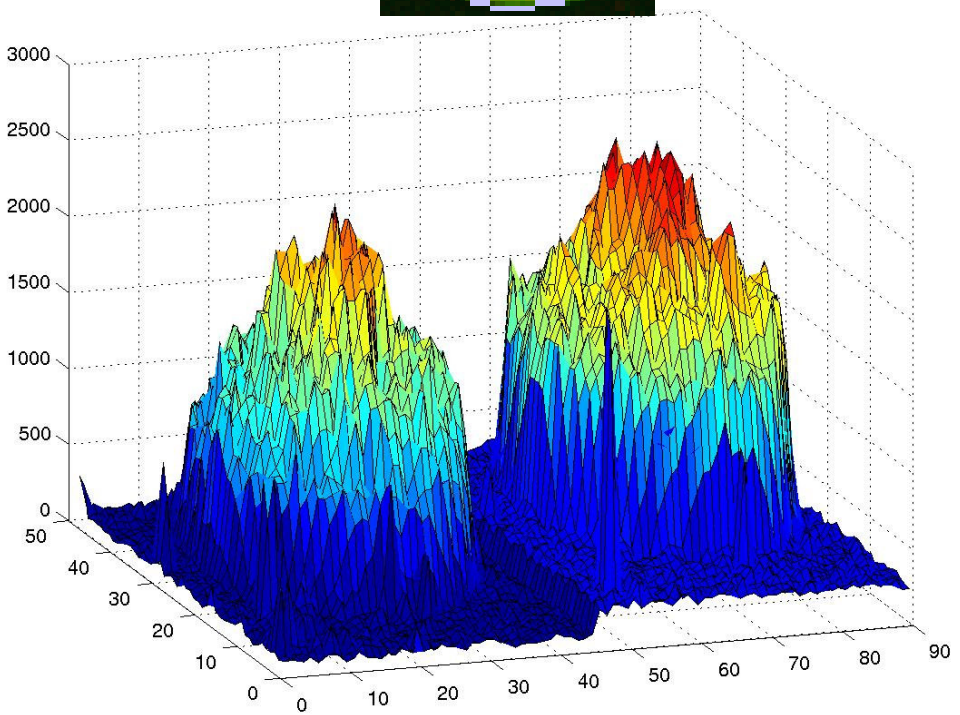


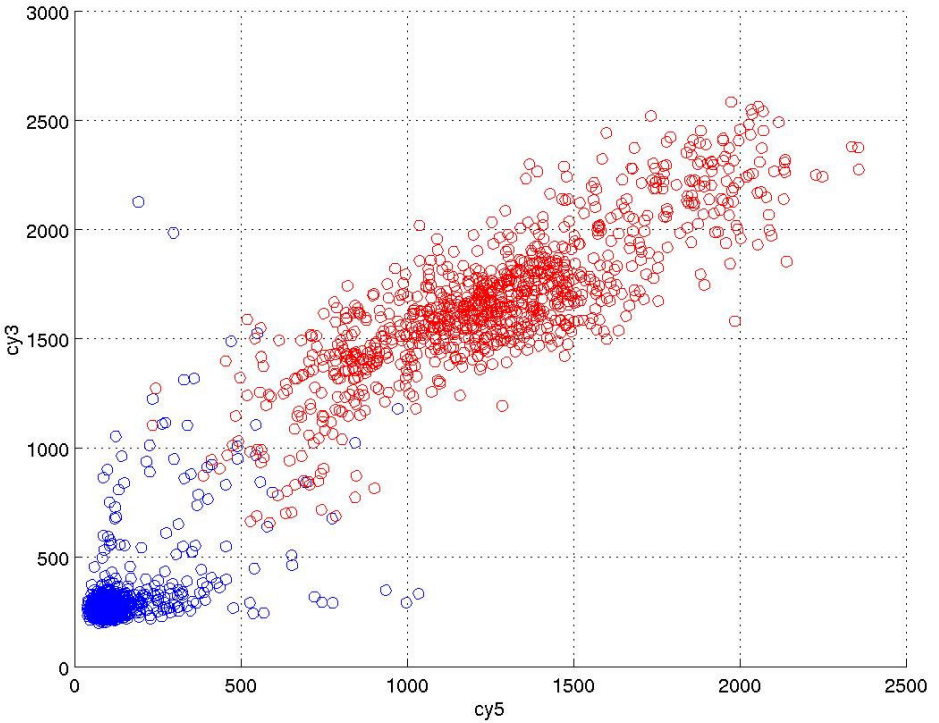
Raw data to the gene expression estimation step

- The raw data of a spot consists on:
  - the pixels values of both channels inside its rectangular region of interest
  - which pixels belong to foreground or backround
- Foreground is the region with spotted cDNA
- Background is the region without it.

Raw data to the gene expression estimation step





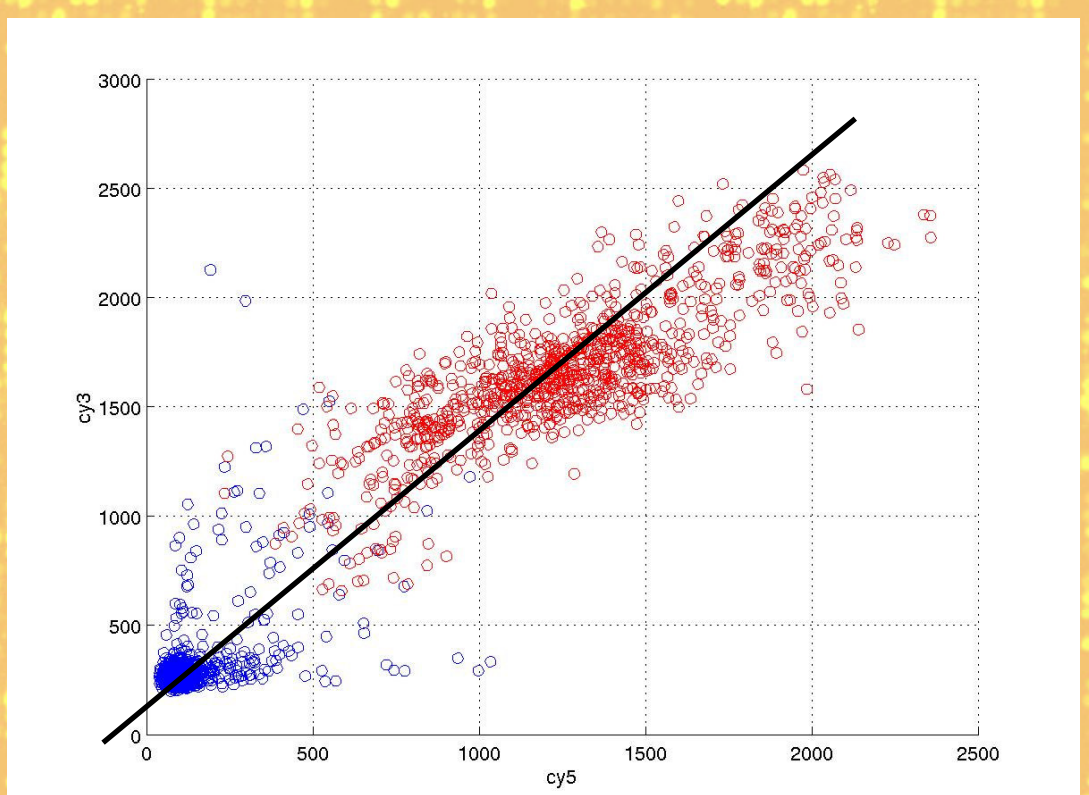


Gene expression estimation

- Is to find a value that represents the relative quantity of mRNA in the two samples.

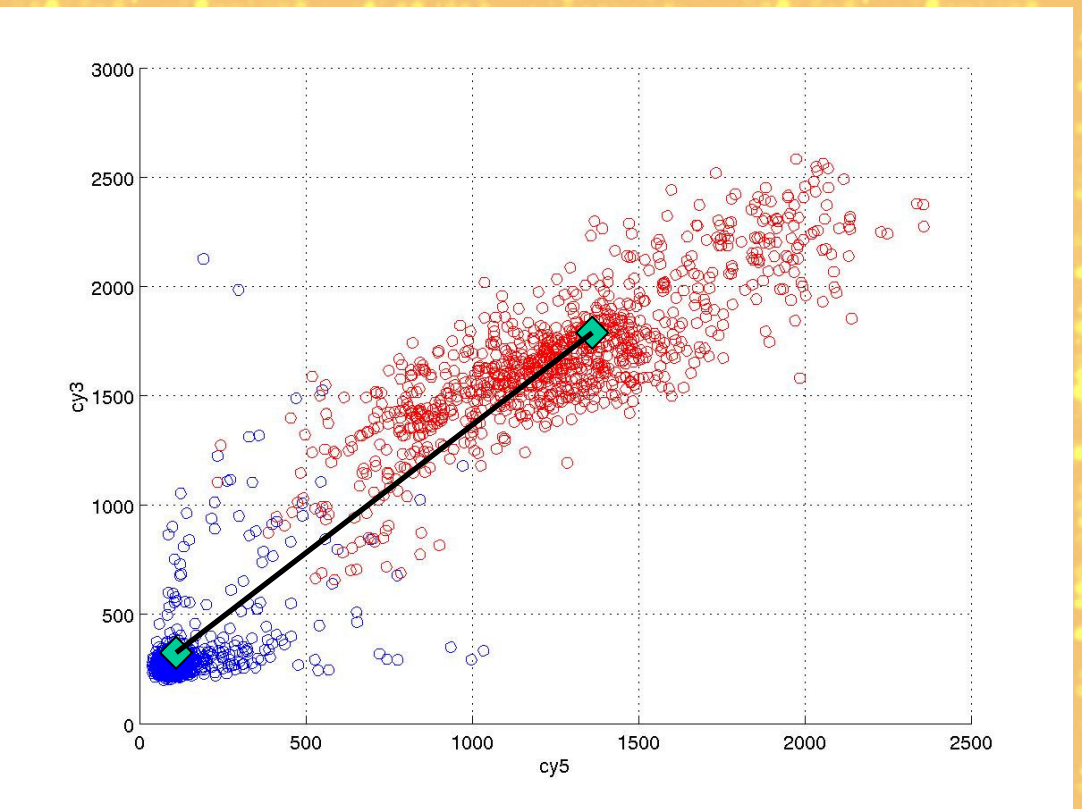
Some techniques to estimate gene expression

- Linear regression or least-squares fit of the values of pixels in the two channels.



Some techniques to estimate gene expression

- (ch1i-ch1b) / (ch2i-ch2b) where chXi is the estimated foreground intensity and chXb is the estimated background intensity of channel X.

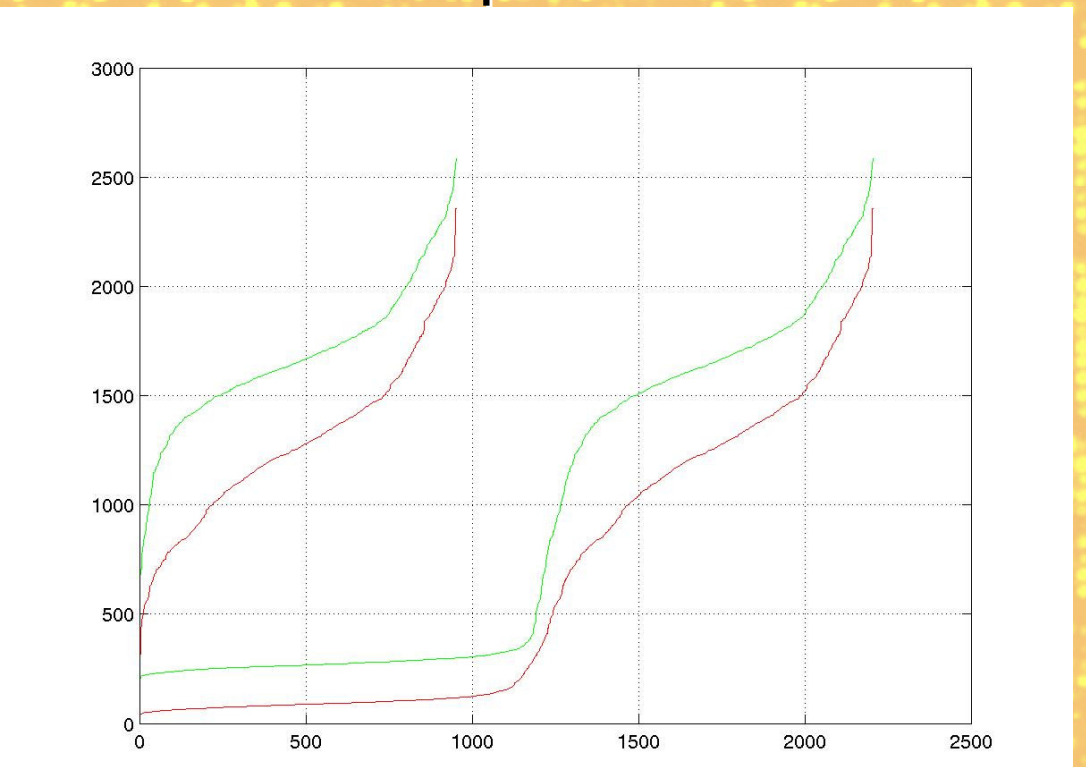


Some techniques to estimate gene expression

- To estimate chXi and chXb we can do:
  - mean or median of all pixels in the foreground and background.
  - mean or median of some percentiles in the foreground and background (fixed region method)
  - mean or median of higher percentiles of all the pixels in the rectangle to estimate chXi and of lower percentiles to estimate the chXb. Foreground and background information is ignored (histogram method)

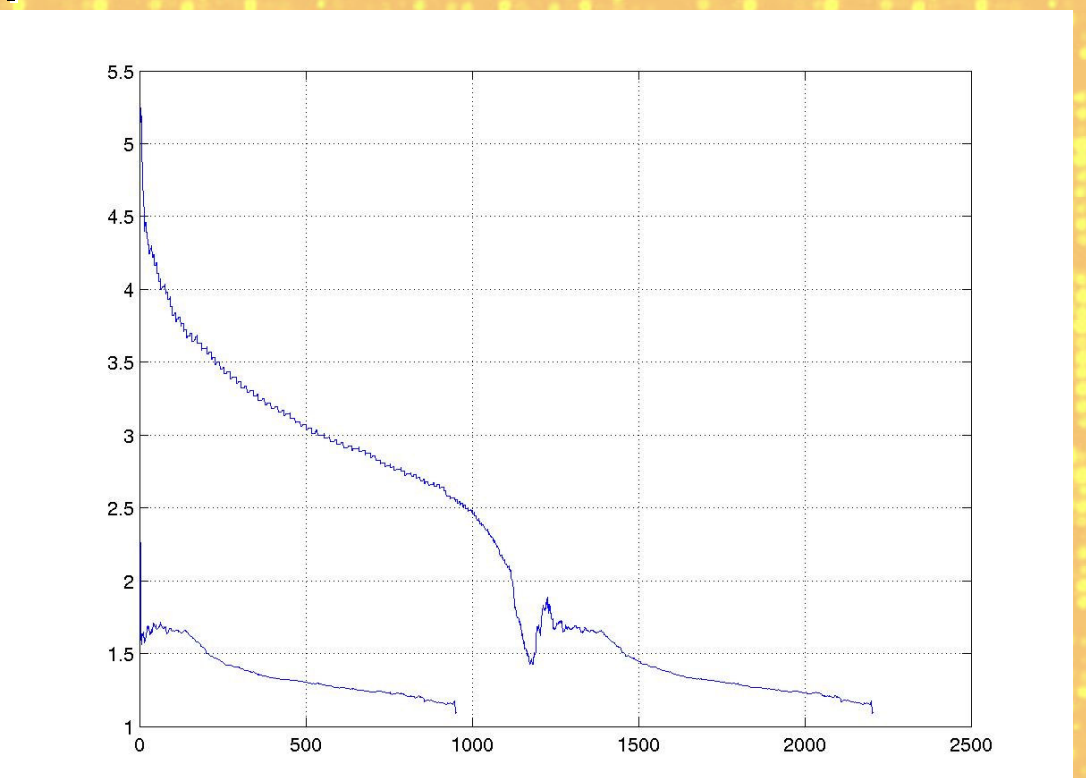
Some techniques to estimate gene expression

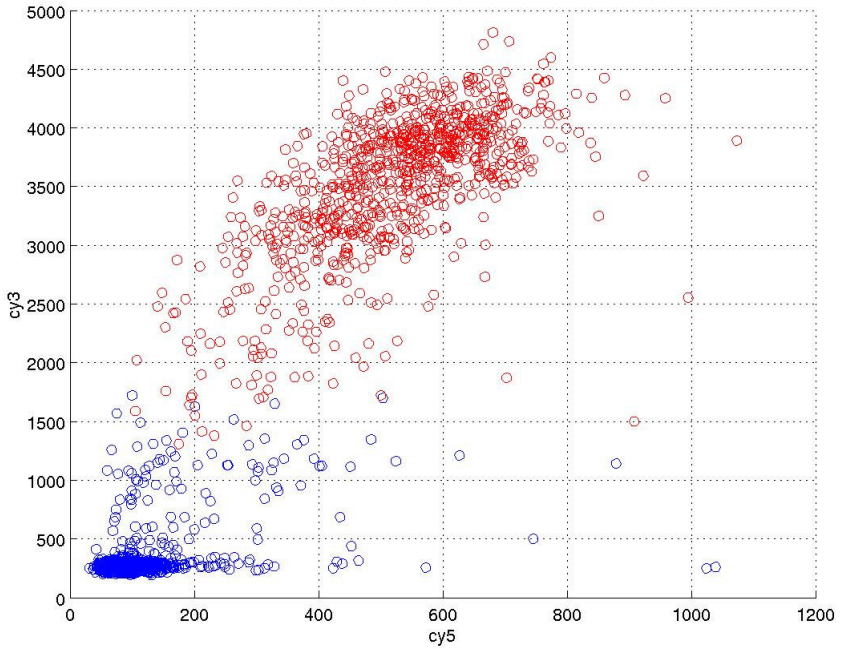
- In both, fixed region and histogram method, we look at parts of graphics like this, with the ordered values of the pixels of both channels.

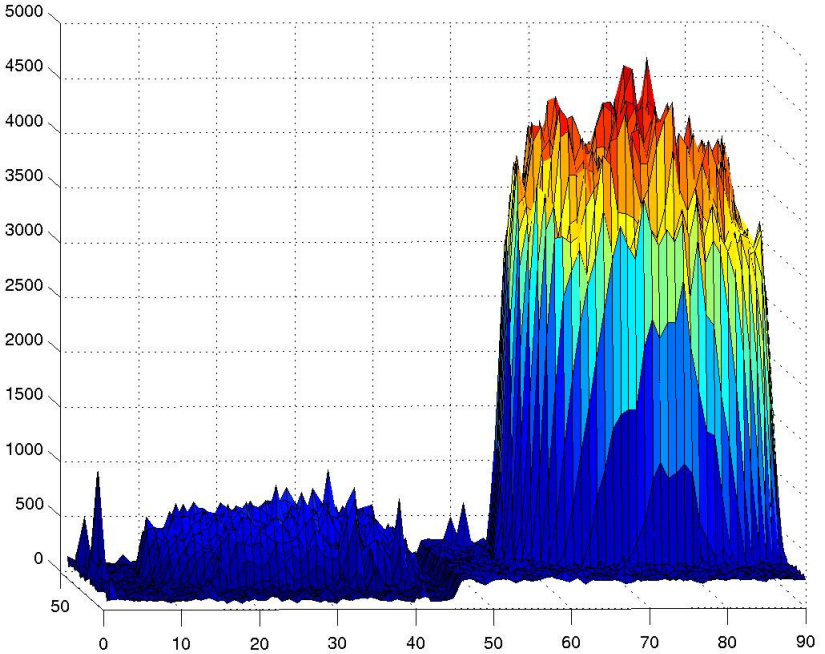


Some techniques to estimate gene expression

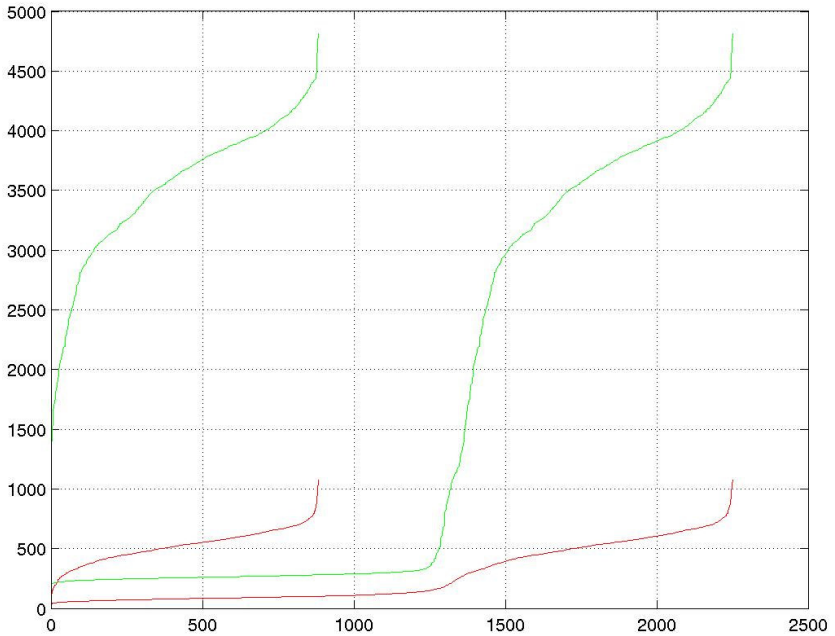
- This graphic shows the quotient green/red, obtained by dividing the curves of the last graphic.

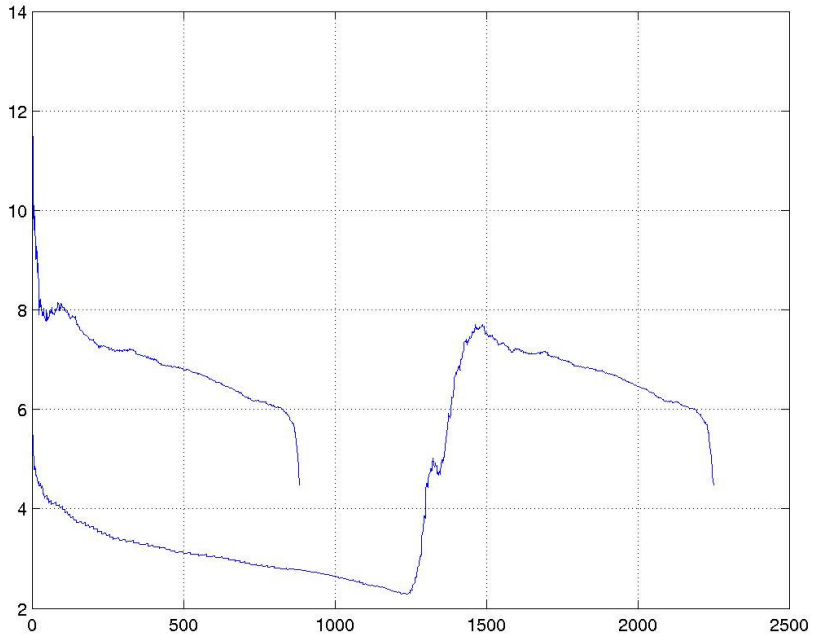


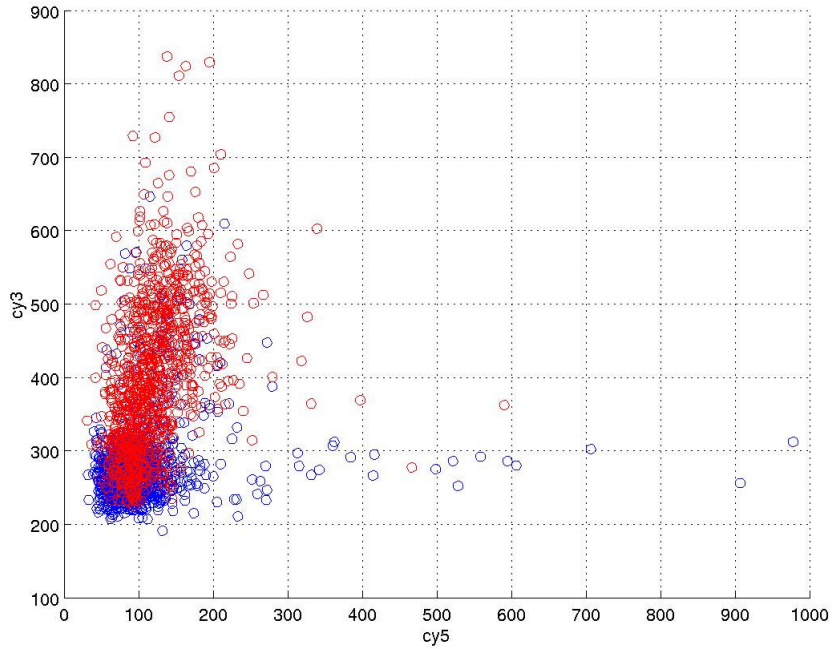

cy3
cy5
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
0
200
400
600
800
1000
1200

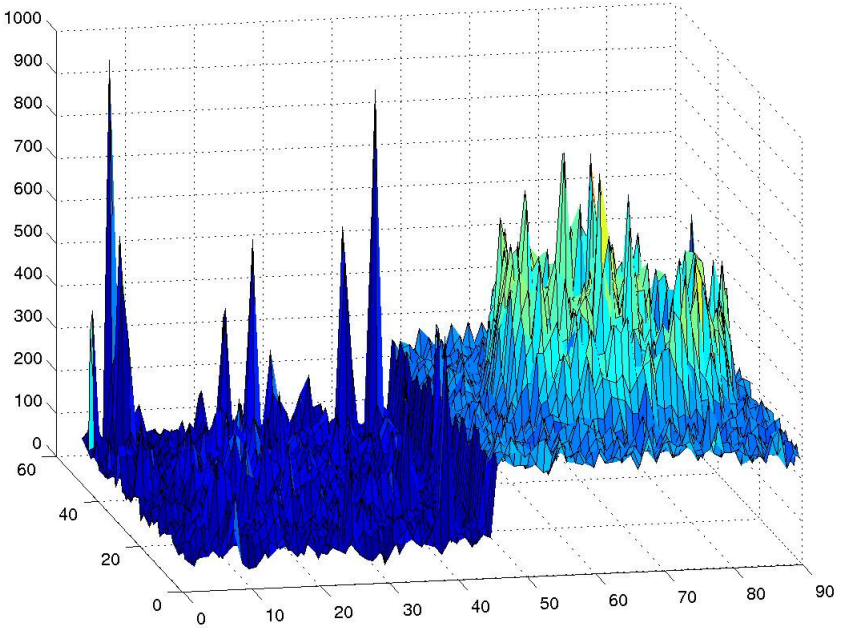

5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
50
0
10
20
30
40
50
60
70
80
90

43A - IL1

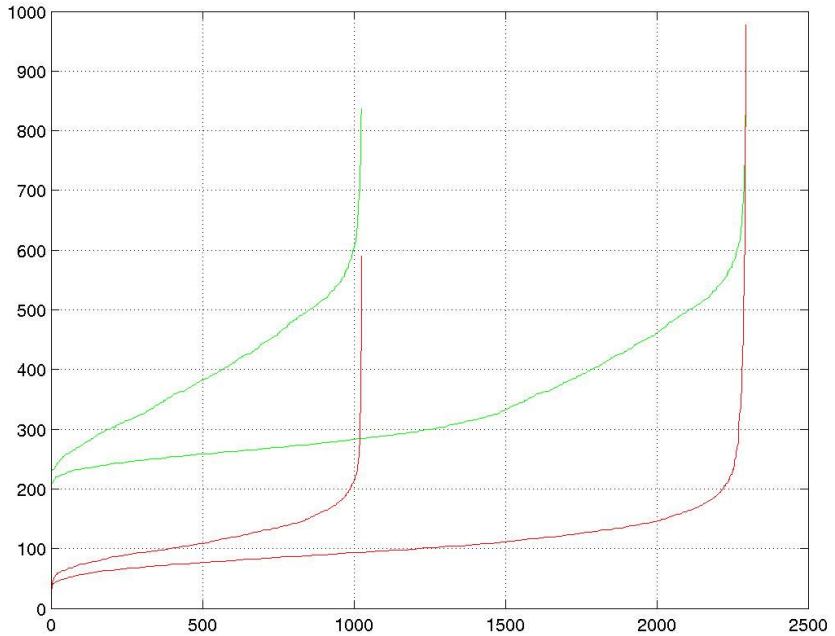

5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
500
1000
1500
2000
2500

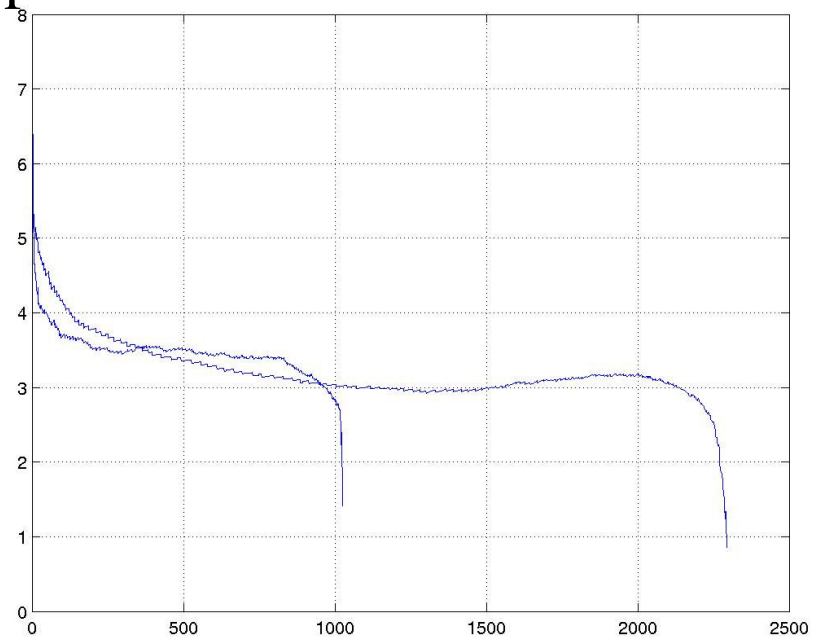

14
12
10
8
6
4
2
0
500
1000
1500
2000
2500

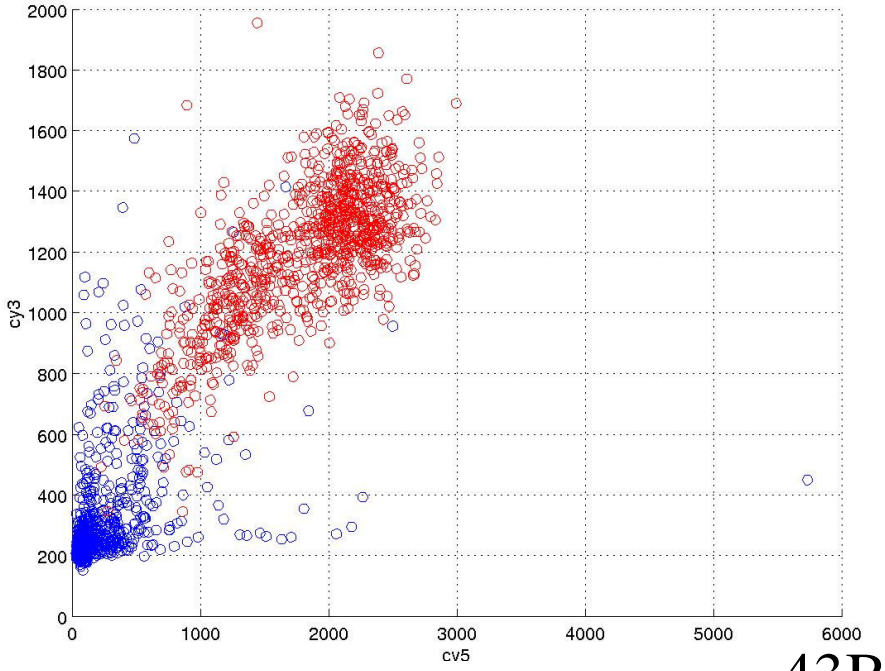


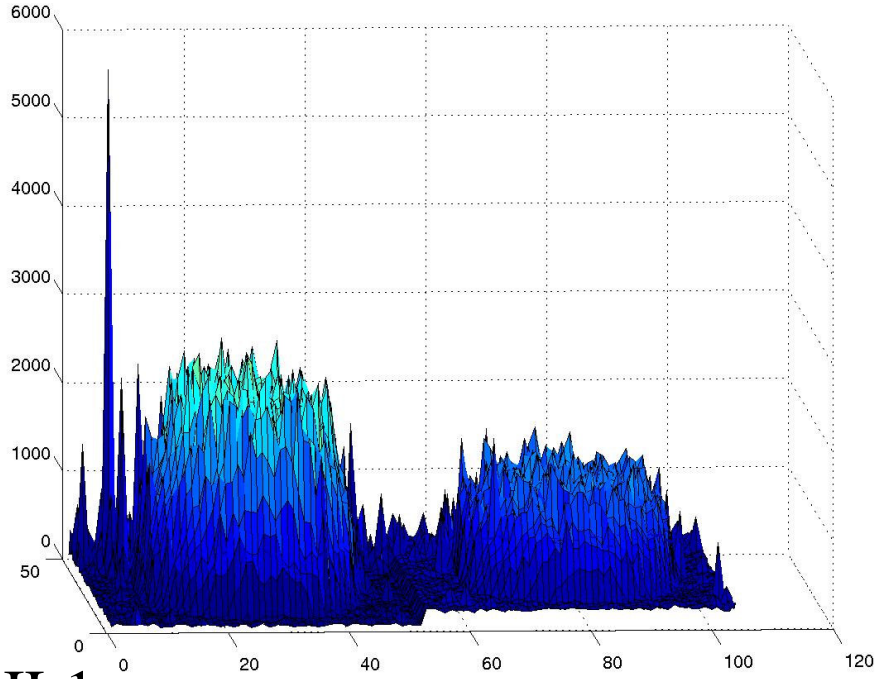


43A - Trp1

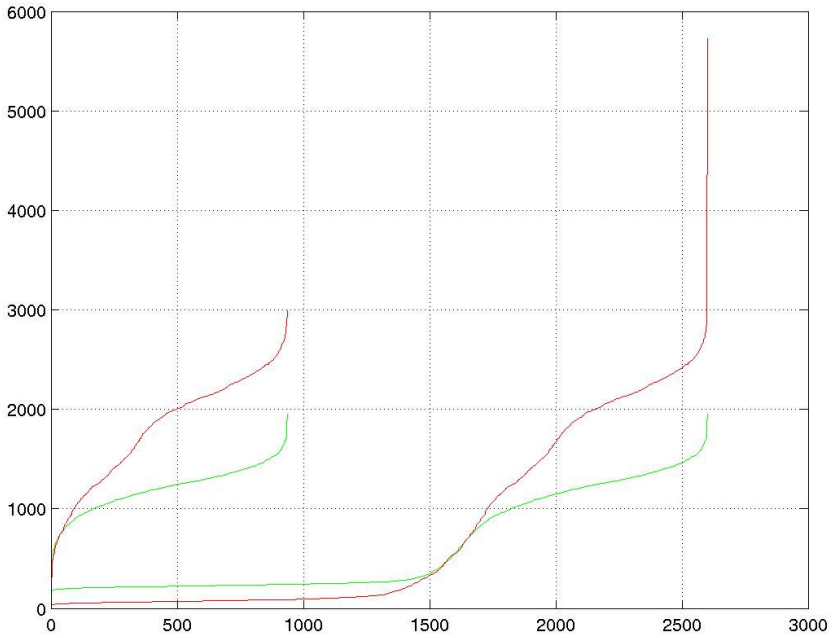


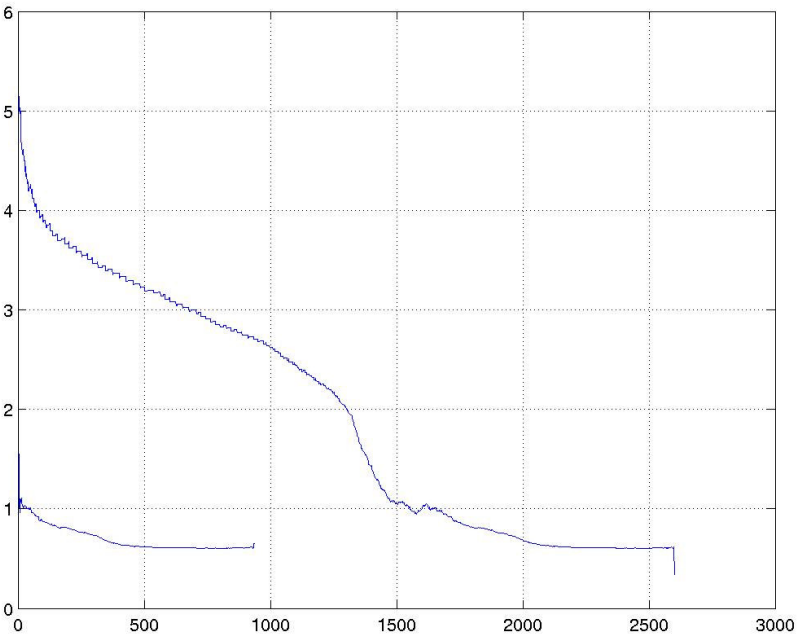



cy3
cv5
0
200
400
600
800
1000
1200
1400
1600
1800
2000
1000
2000
3000
4000
5000
6000


6000
5000
4000
3000
2000
1000
0
50
0
0
20
40
60
80
100
120

43B - IL1


6000
5000
4000
3000
2000
1000
0
0
500
1000
1500
2000
2500
3000


6
5
4
3
2
1
0
0
500
1000
1500
2000
2500
3000

Validation

- We made controlled experiments to test the expression estimation techniques.
- The objective of the experiment was to test how expression was affected by:
  - position in the slide
  - dilution of cDNA
  - length of mRNA fragments
  - being marked with cy3 or cy5

Validation

- We spotted microarrays with 32 blocks, each block with

  6 genes x 5 dilutions x 2 repetitions + 4 landmarks = 64 spots

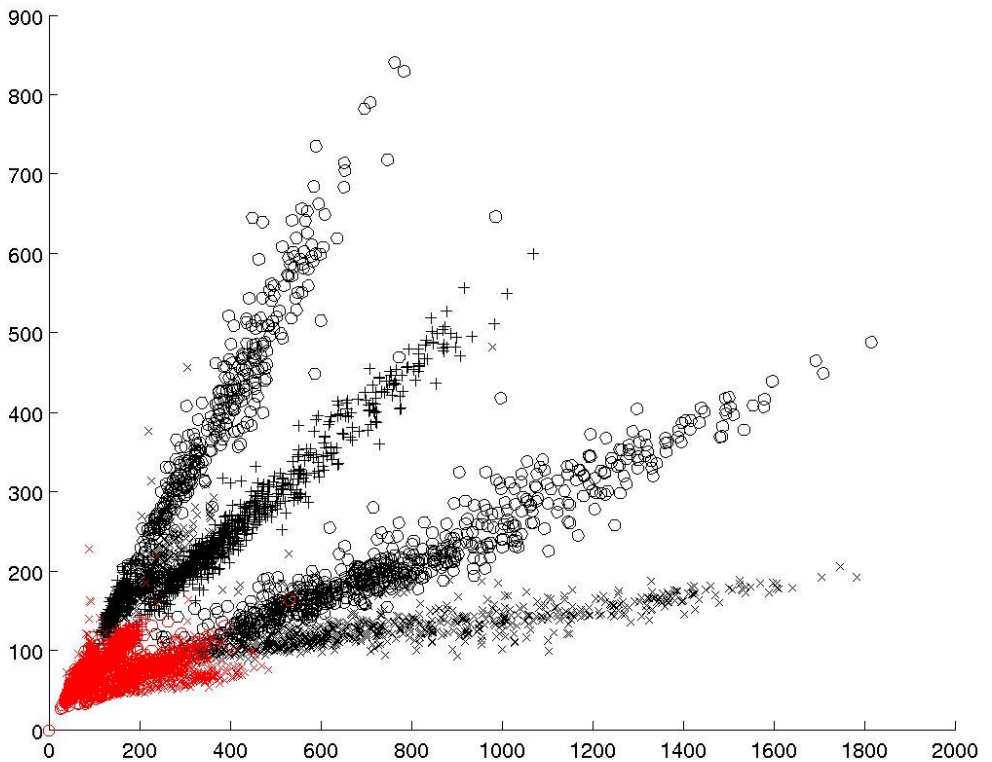
- We made six slides like this and, onto them, we poured six different mRNA soups:

	Dilution					
gene	43A	43B	44A	44B	45A	45B
Irf	1	5	1	2	1	10
Trp	1	5	1	2	1	10
ST0280	1	5	1	2	1	10
IL	5	1	2	1	10	1
Q	5	1	2	1	10	1
Lys	5	1	2	1	10	1

Validation

- Here each point is the value of a spot obtained by the fixed region method. Spots from different dilutions are grouped. The black ones are from the three bigger mRNA fragments, and the red, from the three smaller.

$sqrt((ch1i_A - ch1b_A) \times (ch2i_B - ch2b_B))$

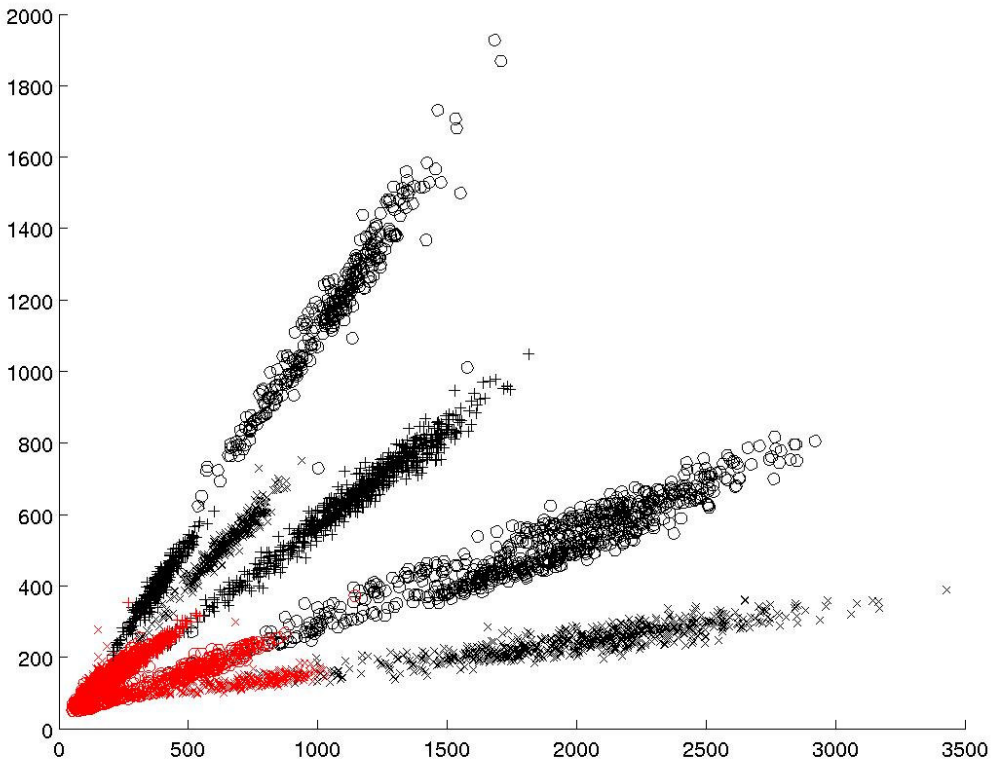


$sqrt((ch2i_A - ch2b_A) \times (ch1i_B - ch1b_B))$

Validation

- And here is the best result, obtained with the histogram method.

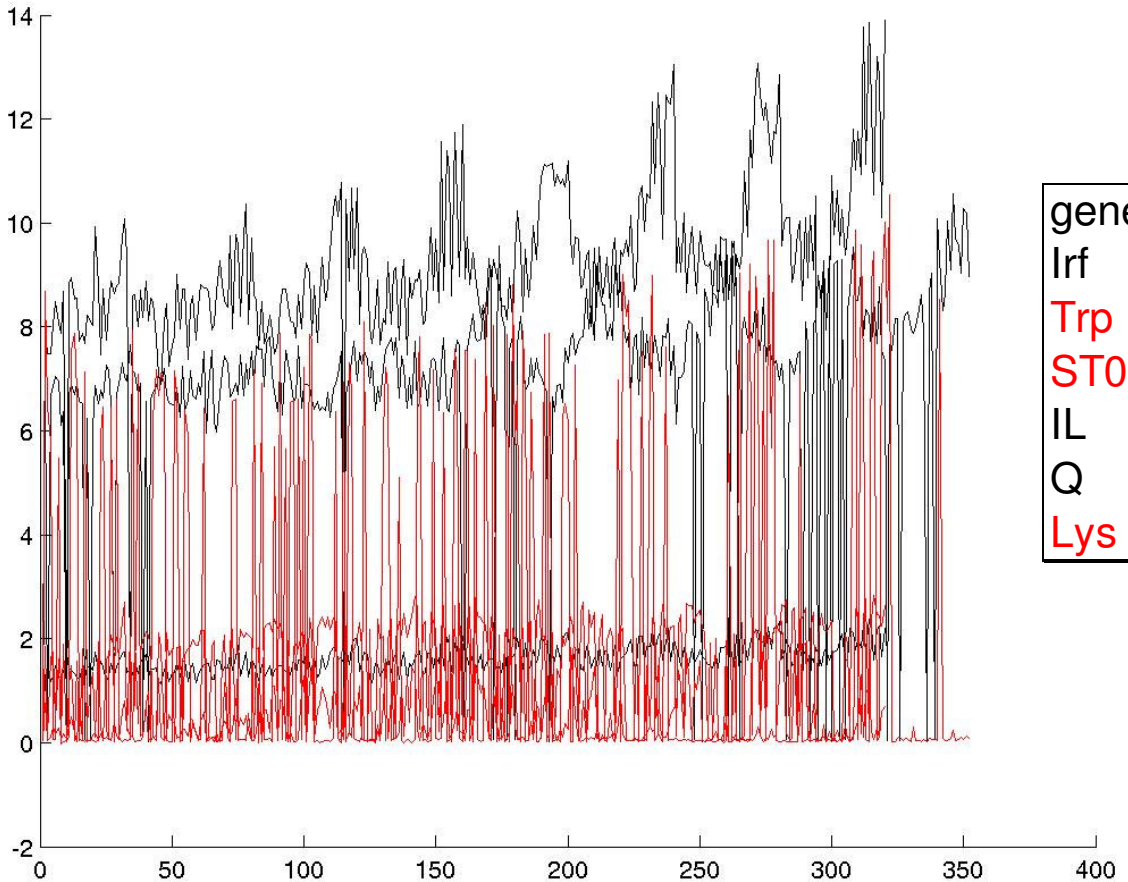
sqrt(($ch1i_A$-$ch1b_A$) x ($ch2i_B$-$ch2b_B$))



sqrt(($ch2i_A$-$ch2b_A$) x ($ch1i_B$-$ch1b_B$))

Validation

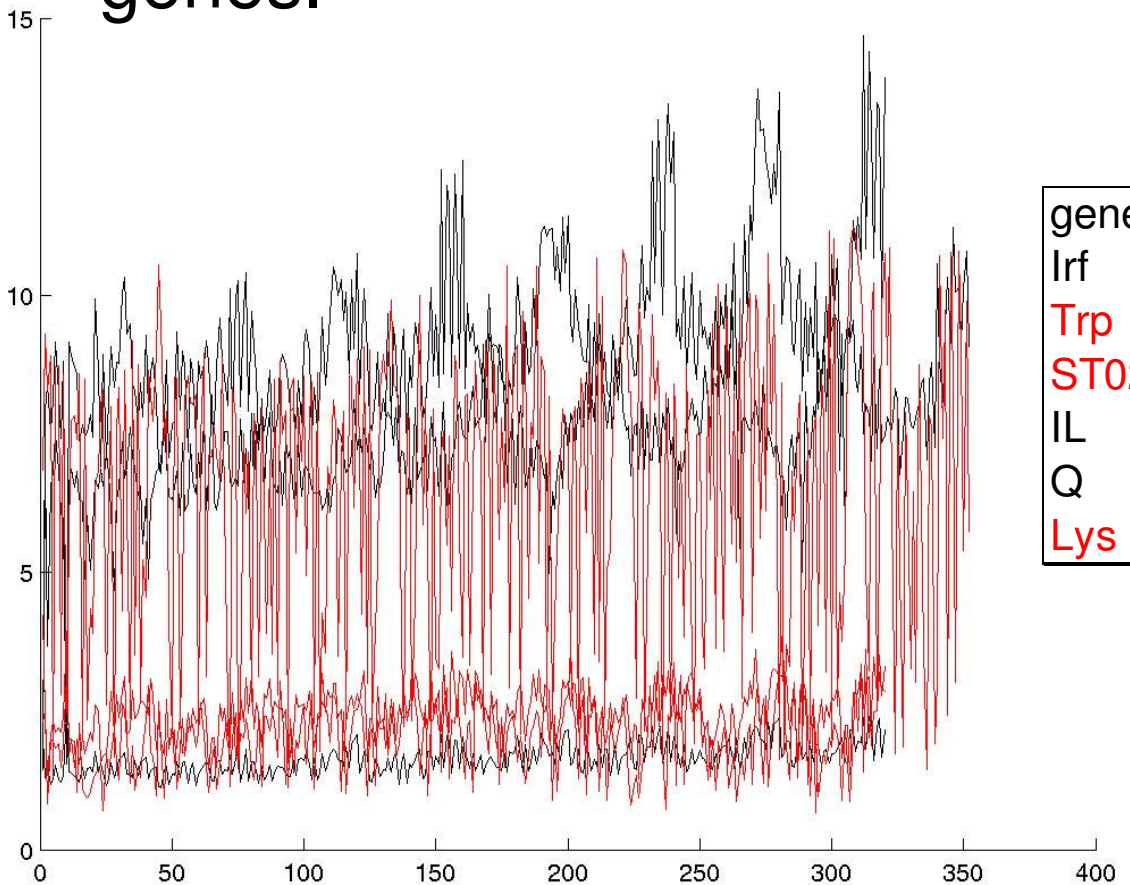
- Applying the least-squares fit to the data of each spot, we obtain results like this for the six genes.



gene	stddev	mean	
Irf	0,2749	1,6176	1,00
Trp	0,4605	0,3999	0,25
ST0280	0,9945	1,4849	0,92
IL	1,8427	9,8712	6,10
Q	2,1836	6,9623	4,30
Lys	3,3600	2,1883	1,35

Validation

- Applying the histogram method to the data of each spot, we obtain results like this for the six genes.



gene	stddev	mean	
Irf	0,2768	1,6411	1,00
Trp	0,6370	2,0420	1,24
ST0280	0,5019	2,4680	1,50
IL	1,5947	9,2869	5,66
Q	1,1552	7,3863	4,50
Lys	2,6532	6,5680	4,00

Normalization

- The expected expression of the gene IRF was 1.0 but the expression found was 1.6
- This is due to the physical properties of the dyes.

Normalization

- When we have a single slide, we must eliminate the constant *k* assuming, when appropriate, that
  - we can normalize all the spots using the expression of a housekeeping gene

Normalization

- When we have a single slide, we must eliminate the constant *k* assuming, when appropriate, that
  - we can normalize all the spots using the expression of a housekeeping gene

$$x = k \frac{(\mathrm{ch1i} - \mathrm{ch1b})}{(\mathrm{ch2i} - \mathrm{ch2b})}$$

Normalization by swap

- Consists on eliminating the influence of the dyes properties by using two slides, and swapping the dye used to label the mRNA sample.
- Use it if you find the single slide normalization hypotheses too strong.

Normalization by swap

- Better results can be achieved by doing swap experiments.

$$x = k\frac{(\mathrm{ch1i_A} - \mathrm{ch1b_A})}{(\mathrm{ch2i_A} - \mathrm{ch2b_A})} = \frac{(\mathrm{ch2i_B} - \mathrm{ch2b_B})}{(\mathrm{ch1i_B} - \mathrm{ch1b_B})} \cdot \frac{1}{k}$$

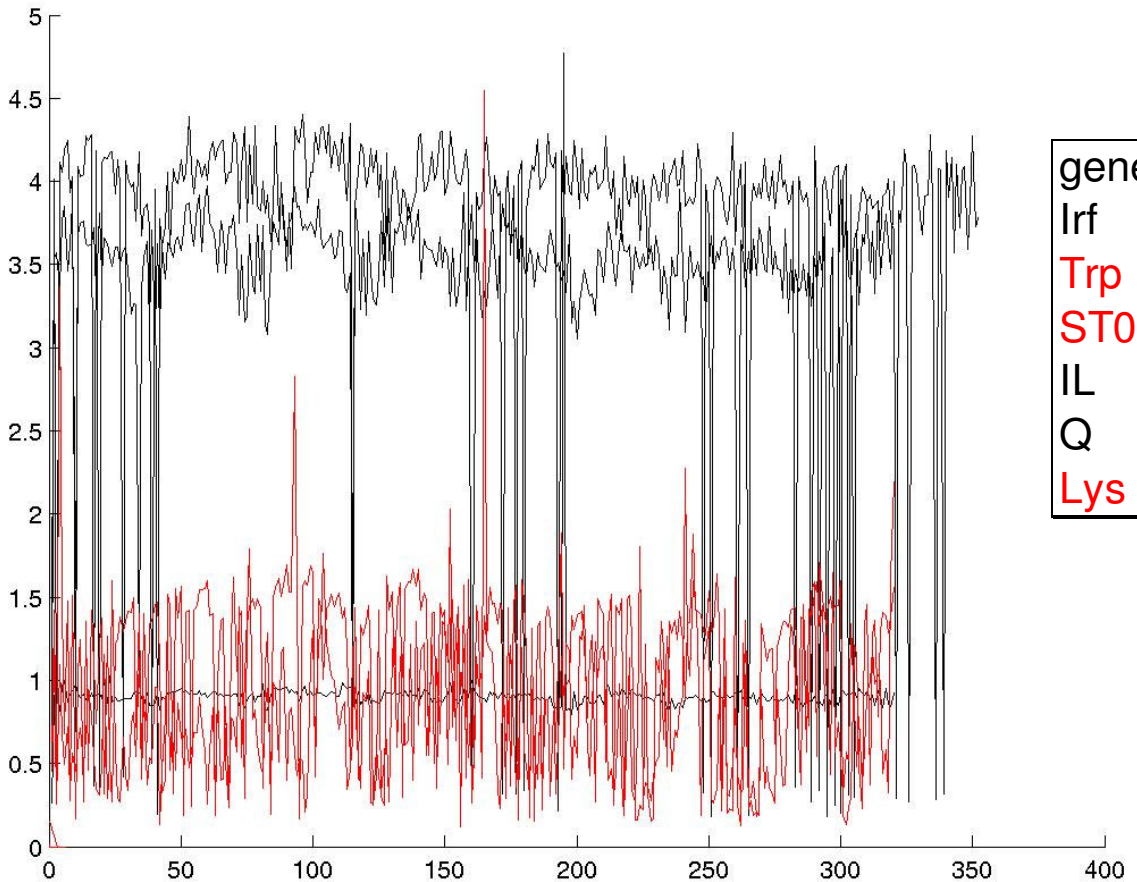
Normalization by swap

- Better results can be achieved by doing swap experiments.

$$x = \sqrt{\frac{(\mathrm{ch1i_A} - \mathrm{ch1b_A})}{(\mathrm{ch2i_A} - \mathrm{ch2b_A})} \cdot \frac{(\mathrm{ch2i_B} - \mathrm{ch2b_B})}{(\mathrm{ch1i_B} - \mathrm{ch1b_B})}}$$

Normalization by swap

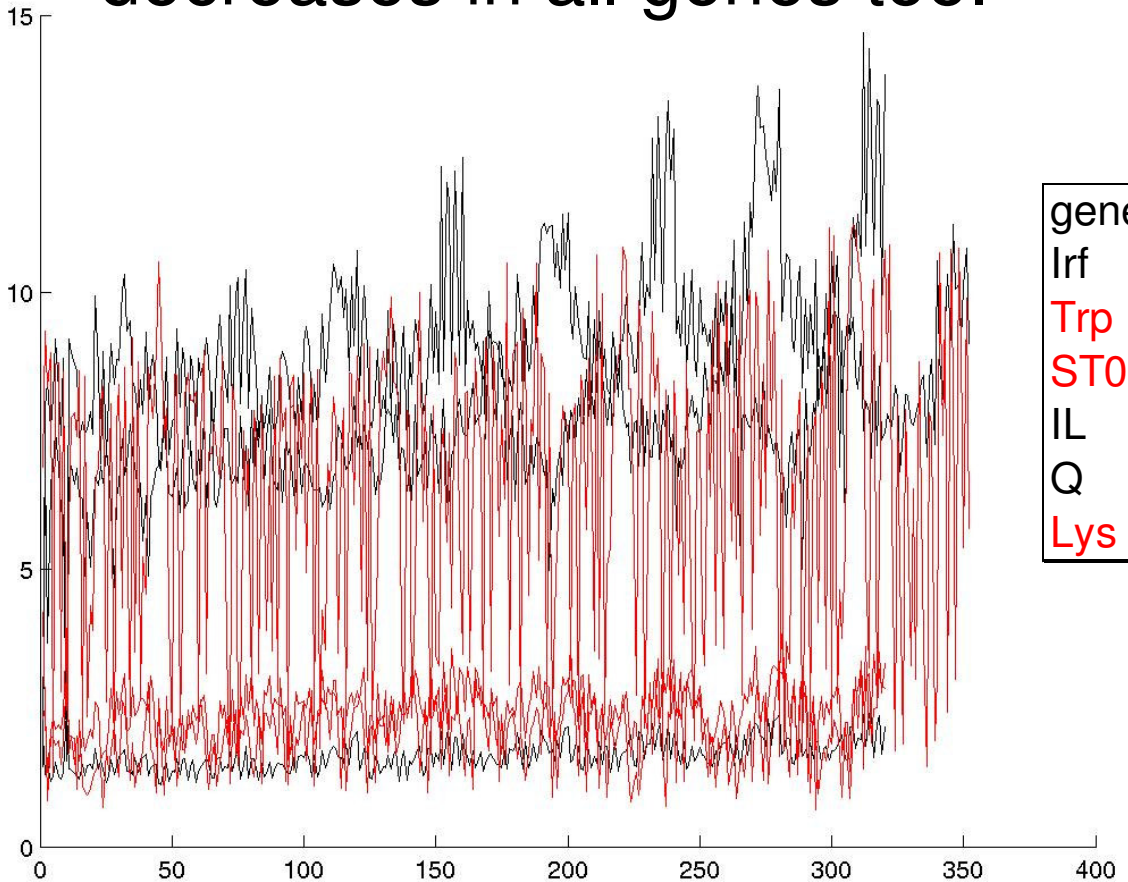
- Using the data obtained by least-sqares fit from the two slides, the deviations decreases in all genes.



gene	stddev	mean	
Irf	0,2224	0,9130	1,00
Trp	0,4039	0,7801	0,85
ST0280	0,5492	1,1251	1,23
IL	0,4567	3,4928	3,83
Q	0,9869	3,7146	4,07
Lys	1,3503	1,2297	1,35

Normalization by swap

- Using the data obtained by the histogram method from the two slides, the deviations decreases in all genes too.



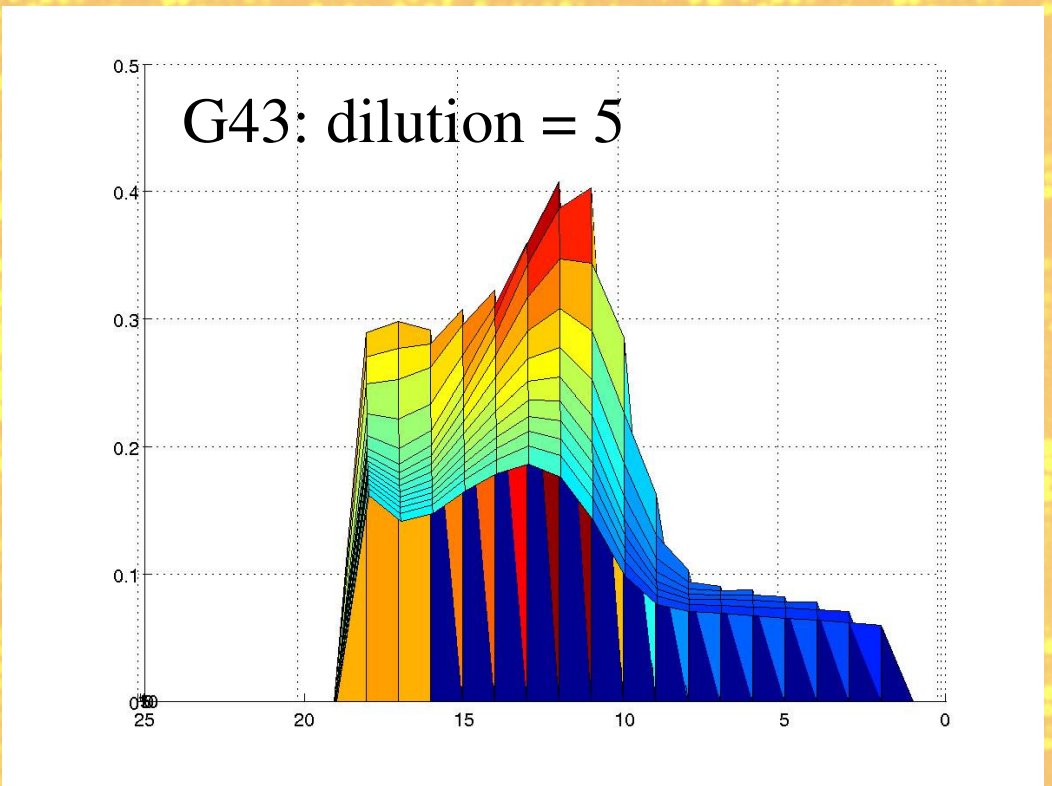
gene	stddev	mean	
Irf	0,0389	0,8959	1,00
Trp	0,1325	1,1096	1,24
ST0280	0,1482	1,3645	1,52
IL	0,2716	3,4475	3,85
Q	0,2964	3,8226	4,27
Lys	0,6194	2,7546	3,07

Normalization by swap

- Assuming that the best estimators are the ones with smaller standard deviation, we analyzed the resulting standard deviation of some different ways of choosing the pixels.

Normalization by swap

- Standard deviations using different values of percentiles for the foreground and bachground. Histogram method.

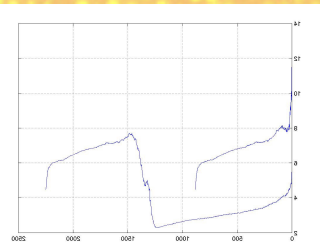


Higher background

Lower background

Higher foreground

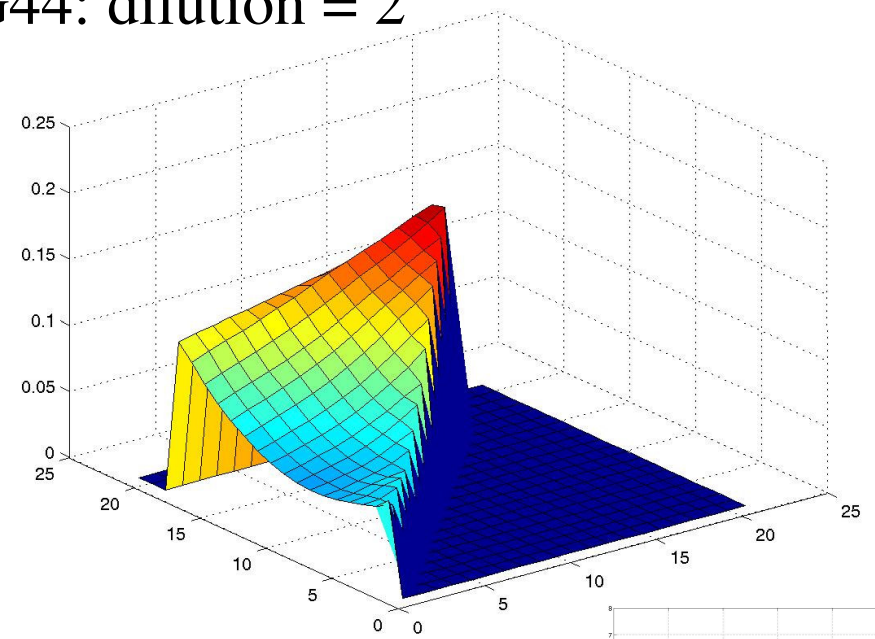
Lower foreground

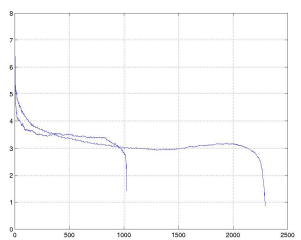


Normalization by swap

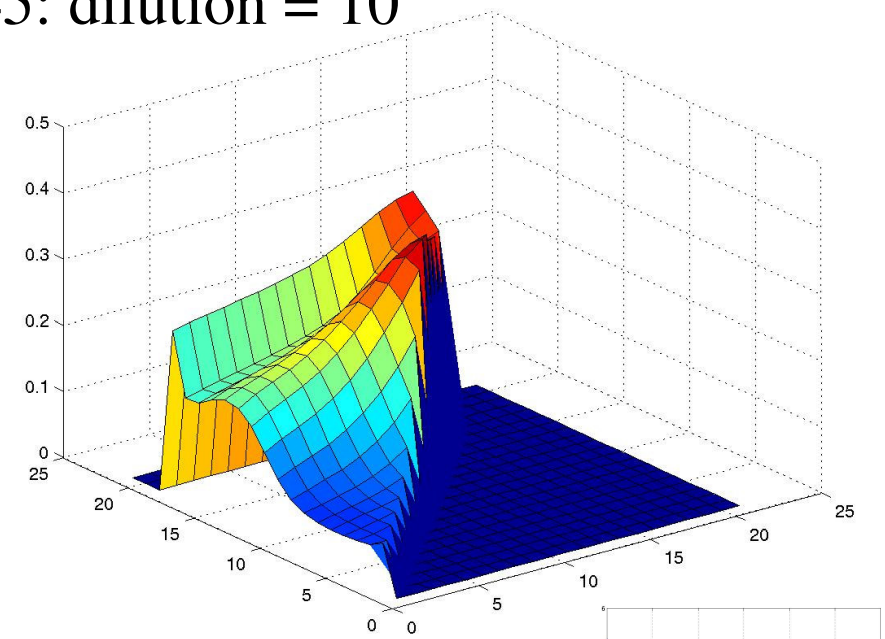
- Standard deviations using different values of percentiles for the foreground and bachground. Histogram method.

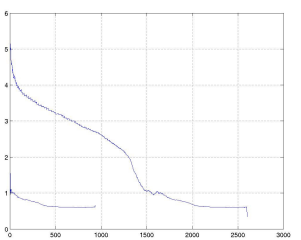
G44: dilution = 2





G45: dilution = 10





Gene expression generation

```
HEADER      SPOT  GRID  TOP   LEFT  BOT   RIGHT ROW   COL   CH1I  CH1B  CH1AB CH2I  CH2B  CH2AB SPIX  BGPIX EDGE  RAT2  MRAT  REGR
      CORR  LFRAT CH1GTB1     CH2GTB1     CH1GTB1     CH2GTB2     CH1EDGEA    CH2EDGEA    FLAG  CH1KSD      CH1KSP      CH2KSD
      CH2KSP
REMARK
REMARK
REMARK
REMARK
REMARK      DATE  21-Mar-2002
REMARK      TIME  16:51:57
SPOT  1     1     88    24    102   40    1     1     7341  6704  6818  4669  4016  4312  105   150   0     0     0.9624      0.636
      1.0000      0.0000      0.8952      0.9143        0   0.01905     -8.68E+03   -3.30E+03   0         0   0.00E+00        0   0.00E+00
SPOT  2     1     88    40    102   53    1     2     7075  6704  6708  4419  3920  3989  89    121   0     0     0.9751      0.5965
      1.0000      0.0000                                                                                    E+00        0   0.00E+00
SPOT  3     1     88                                                                                        0.6031      0.5906
      1.0000      0.0000                                                                                       0   0.00E+00
SPOT  4     1     88                                                                                        1.422 0.7494
      1.0000      0.0000                                                                                       0   0.00E+00
SPOT  5     1     88                                                                                        1.018 0.8077
      1.0000      0.0000                                                                                    0.00E+00
SPOT  6     1     88                                                                                        0.9687      0.6442
      1.0000      0.0000                                                                                       0   0.00E+00
SPOT  7     1     88                                                                                        1.125 0.6169
      1.0000      0.0000                                                                                       0   0.00E+00
SPOT  8     1     88                                                                                        1.044 0.6949
      1.0000      0.0000                                                                                    0.00E+00
SPOT  9     1     88                                                                                        0.7028      0.7353
      1.0000      0.0000                                                                                       0   0.00E+00
SPOT  10    1     88                                                                                        0.4444      0.4603
      1.0000      0.0000                                                                                    E+00        0   0.00E+00
SPOT  11    1     102                                                                                       1.417 0.3783
      1.0000      0.0000                                                                                    0.00E+00
SPOT  12    1     102   40    118   53    2     2     10875 7552  7803  7889  5232  5464  75    163   0     0     0.7956      0.8394
      1.0000      0.0000        1     1   0.2533      0.5467      -3.89E+04   -9.60E+04   0         0   0.00E+00        0   0.00E+00
SPOT  13    1     102   53    118   68    2     3     9546  7136  7245  6922  4688  4759  98    174   0     0     0.9084      0.8939
      1.0000      0.0000        1   0.9898      0.1327      0.4286      1.48E+04    -1.99E+04   0         0   0.00E+00        0   0.00E+00
SPOT  14    1     102   68    118   83    2     4     49743 9808  12869 26888 7288  8971  100   172   0     0     0.5192      0.6069
      1.0000      0.0000        1     1     1   0.99  -2.61E+05   -3.82E+05   0         0   0.00E+00        0   0.00E+00
SPOT  15    1     102   83    118   97    2     5     11367 8160  9020  7880  6032  6499  92    163   0     0     0.5999      0.6874
      1.0000      0.0000        1   0.9022      0.2609      0.1739      5.17E+04    1.76E+04    0         0   0.00E+00        0   0.00E+00
```

- The program saves the expression data in a tab separated text file
- The file has the same format of the ones generated by *ScanAlyze*

Conclusion

- We created an automatic method for segmenting microarray images and estimating gene expression.
- The process was validated by controlled biochemical experiments.
- Some future steps:
  - Automatic tilt correction
  - Automatic identification of bad spots
  - Statistically test if the controlled experiments represent properly real experiments.
  - Automatic choice of the best estimation method
  - Assign error bars to expression