

## PreP: Gene-Expression Data Pre-Processing Tool

Jorge García de la Nava<sup>1</sup>, Sacha van Hijum<sup>2</sup> and Oswaldo Trelles<sup>1,\*</sup>

<sup>1</sup>Computer Architecture Department, Universidad de Málaga, 29080, Málaga, Spain

<sup>2</sup>University of Groningen, Molecular Genetics, The Netherlands.

Gene Expression data analysis procedures are strongly dependent on the quality of the measured data. The application of data pre-processing methodologies can lead to a better definition of clusters, class markers, etc. producing a significant improvement in results

There are many sources of systematic variation in microarray experiments which affect the measured gene expression level. Normalization is the term most frequently used in the literature to describe noise-signal ratio improvements. However, strictly speaking, normalization refers to transform a set of values to have on average zero and standard deviation equal to 1. Therefore, along this document we will use the term Pre-Processing to describe the process of removing such systematic variations

This manual is a work in progress  
by Jorge García de la Nava  
[gdl@ac.uma.es](mailto:gdl@ac.uma.es)

Biological side is covered by  
Sacha van Hijum  
[S.A.F.T.van.Hijum@biol.rug.nl](mailto:S.A.F.T.van.Hijum@biol.rug.nl)

Project Director  
Dr. Oswaldo Trelles  
[ots@ac.uma.es](mailto:ots@ac.uma.es)

Computer Architecture Department  
<http://www.ac.uma.es>  
University of Málaga, Spain

Bioinformatics and Genomics Laboratory, University of Málaga  
<http://chirimoyo.ac.uma.es/bglab/index.html>

**Express-Fingerprints**  
Expression profiles as fingerprints  
for the safety evaluation of new strains,  
including GMOs used in bioprocessed food

### ***Consortium***

Génétique Microbienne, INRA, France  
University of Groningen, Molecular Genetics, The Netherlands.  
Mathématique, Informatique et Génome (MIG), INRA, France  
Genetics and Microbiology, Chr. Hansen A/S, Denmark.  
Vitavaleur, DANONE Vitapole, France.  
Instytut Biochemii i Biofizyki PAN, Poland  
Computer Architecture Department., University of Malaga, Spain



## Contents

1. Introduction.....	5
1.1 Motivation	
<b>1.2 Objective</b>	
<b>1.3 The initial source of data</b>	
<b>1.4 Data Pre-Processing</b>	
1.5 Post-processing Procedures	
2 Dictionary of common terms.....	8
3 Organization of a PreP project.....	10
3.1 Working at the Project level	
3.2 The Load Step	
3.3 Loading slides and the Screen Layout	
3.4 Slide data visualization	
3.5 Available functionality for each label	
3.6 Assigning function to labels	
3.7 Ending the load Step: the <i>slide</i> structure	
4 Slides Visualization.....	19
4.1 Screen Layout	
4.2 Viewing types	
4.2.1 <i>Slide view</i>	
4.2.2 Coherent Slide view	
4.2.3 Slide view with Quality	
4.2.4 AM Graph	
4.2.5 AM Graph by blocks	
4.2.6 RG Graph	
4.2.7 Box Graph	
4.2.8 Values Density	
4.2.9 By block values density	
4.2.10 Intensity-Intensity Graph	
4.2.11 Scatter plot of replicates	
4.2.12 Standard deviation of replicates	
4.2.13 Replicates Correlation	
4.2.14 Normality of Replicates	
4.3 Visualization Controls	
4.3.1 Slide Channels	
4.3.2 <i>Zoom</i>	
5 Selection, groups and Slide marks.....	33
5.1 Slides selection	
5.2 Grouping <i>slides</i>	
5.3 Setting marks to <i>slides</i>	
6 Pre-Processing.....	35



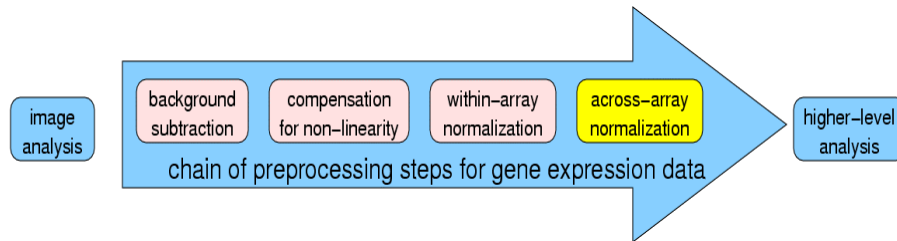
- 6.1 *Lowess* adjust
- 6.2 Selection of the block size
- 6.3 Set Thresholds
- 6.4 Grouping replicate points
- 6.5 Double *scan* regression
  
- 7 Steps: Operation over the states..... 40
  - 7.1 Ratio correction
  - 7.2 Scale Adjust
  - 7.3 Filtering
  - 7.4 Solving *dye-swap*
  - 7.5 Solving replications
  - 7.6 Solving double-scan
  
- 8 Annexes
  - 8.1 Annexe A, File format for *slide* data..... 44
  - 8.2 Annexe B, Output File Format (*engine* format). 45
  - 8.3 Annexe C, The initial source of data..... 47
  - 8.4 Annexe D: Image Analysis issues..... 49

**Málaga, February 24<sup>th</sup>, 2003**



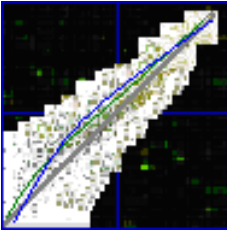
## The Pre-Processing chain

Normalization involves a diverse gallery of methods with different but combined scope. At the end (or between-array normalization) we devise **engene** that should take place after image analysis and before high level analysis, including visualization (for human analysis), clustering, classification etc. Next picture shows a sketch of this chain as we see it.



*Sketch of the chain of pre-processing steps that should be applied to gene expression data.*

Most image analysis programs estimate the level of additive background noise by analysing the signal intensities around the spots. Usually, the estimated background intensity is subtracted from the intensities measured for the spots themselves. The intensities may still not be linearly related to the numbers of corresponding mRNA molecules, for example due to saturation effects (we are currently working on this aspects by a double-scanning approach).



# PreP: Gene-Expression Data Pre-Processing Tool

Jorge García de la Nava<sup>1</sup>, Sacha van Hijum<sup>2</sup> and Oswaldo Trelles<sup>1,\*</sup>

<sup>1</sup>Computer Architecture Department, Universidad de Málaga, 29080, Málaga, Spain

<sup>2</sup>University of Groningen, Molecular Genetics, The Netherlands.

## User Manual

### 1.- Introduction.

#### 1.1 Motivation

Genomics aims to decipher the complete catalogue of genes for the myriad of living organisms, however, in the post-genomic era the interest is moving towards a dynamical view of the organism: the *proteome*, or the set of protein expressed in a given cell. This is to say, rather than knowing the set of genes available in a given organism, the interest is focused in understanding the genes expressed under certain environmental circumstances, and specially, how this pattern changes when the given conditions are modified.

The regulation of cellular processes is achieved by modifying the relative proportions of proteins present in the cell, therefore, the key element in proteome analysis is identify these variations. Unfortunately, proteins are hard to analyse due to technical limitations and intrinsic protein characteristics. However, it is assumed that the level of transcripts (mRNA) is linearly related to the level of proteins they encode; and mRNA transcripts are much easier to obtain. Thus, measuring the concentration of mRNA molecules it is possible to infer the concentration of the distinct protein molecule in a sample of cells

#### 1.2 Objective

Data normalization procedures, aim to identify and remove sources of systematic variation in the measured data (fluorescence intensities), other than differential expression

There are several steps along the process of measurement DNA-arrays, but none of these steps have an optimal and strictly defined procedure since this procedure is dependent of the final objective of the experiment. However, for each step the objective is to minimize the error ratio in the measure, and several specific techniques has been developed to this aim. This tool aims to help in the data pre-processing task, taking advantage of the different techniques available to reduce the error in the data, and also by introducing a new approach based on a *double-scanning* of the fluorescence signals.

#### 1.3 The initial source of data

PreP application works with data obtained from DNA-array experiments.



In brief, a microarray gene expression experiment works as follows: DNA sequences relevant to the biological question are selected and printed onto glass slides, one 'spot' for each sequence in a rectangular grid layout. These target sequences are linked to the glass surface. From the biological sample in question as well as from a reference sample, mRNA is extracted from which fluorescent labelled DNA is produced by reverse transcription using labelled nucleotides. The labelled DNA sequences are brought onto the array surface for hybridisation. The reference and probe samples are labelled with different dyes and hybridised to the array at the same time. Separate fluorescence images of the reference and probe dyes are taken using a confocal laser scanning microscope. The ratios of reference and probe dye intensities at each spot are interpreted as changes of gene expression levels between the two samples.

Traditionally, after completion of the hybridisation processes the slide glass is scanned through two channels for Cy3 and Cy5 to obtain the fluorescence signals, thus producing two image files from one slide glass: one from the red channel and a second image from the green channel. A numeric representation of the spot intensities is obtained by image processing procedures and tabulated as a matrix, each row representing a given spot (see Annex C for details). This will be the input to the pre-processing pipeline.

## 1.4 Data Pre-Processing

There are many sources of systematic variation in microarray experiments which affect the measured gene expression level. **PreP** is a software tool aimed to supply adequate procedures to remove such variation.

Data variability is often introduced as a mixture of some of the following concepts:

- sample preparation: fluctuation in efficiency of reverse transcription and labelling
- differences in labelling efficiency between the two fluorescence dyes
- background interference: noise problems with detecting low expression levels
- intensity and spatial dependence in measurements.
- variability within and between slides
  - amount of DNA spotted (pin geometry, fluctuations in volume spotted, fixation to slide)
  - efficiency of hybridization, differences in spot arrays

**PreP** supplies different procedures aimed to remove the noise. Some of them can be applied in any context (such as the normalization –adjusting and ratio scaling). Other techniques require some specific conditions (e.g. gene replication). In any case, once the error has been removed as much as possible, the quantity of interest for each spot is the ratio of the amounts of label dye in the two channels, the target/control ratio(\*).

(\*) Note: In the literature there exist at least two confusing nomenclature systems for referring to hybridization partners. Both use common terms: "probes" and "targets". According to the nomenclature recommended by B. Phimister of Nature Genetics, a "probe" is the tethered nucleic acid with known sequence,



whereas a "target" is the free nucleic acid sample whose identity/abundance is being detected. This document follows that recommendation. See Nature Genetics volume 21 supplement pp 1 - 60, 1999.

## 1.5 Post-processing Procedures

The final objective of gene-expression experiments is focused in understanding how genes express under certain environmental circumstances, and specially, how this pattern changes when the given conditions are modified. The two key applications for processing gene-expression data collections are clustering and classification. The former aims to identify genes with similar expression patterns from which their involvement in related biological processes may be deduced; while the latter places an unknown object (gene or experiment) in one and only one of the *a priori* defined groups. Clustering and classification procedures are strongly dependent on the quality of the gene-expression data. In both cases the application of data pre-processing methodologies can lead to a better definition of clusters, producing a significant improvement in results.



## 2.- Glossary of common terms

**Project** A **PreP** project is a collection of states. Each state is the result of applying a given process over the previous state. Each state is self-contained, that is to say, it contains all the necessary information to produce a new state.

**Current state (last or active state).** The last state is the *current state*, that is to say, the state over which the procedures are applied (all the rest conform the “history”).

**State:** Each state is composed of a collection of *slides*. The *slides* represent and contain the information obtained by *scanning* a given DNA chip. The *slide* have associated a name and, when necessary, a set of pre-computed values to be used in a new step. In general, the slide-name resemble the experimental conditions.

**Slide:** A *slide* is a collection of *spots*. Each spot have a set of values that correspond to luminosity intensities, position in the chip, labels, etc. referred to a given DNA molecule.

**Replication** (replicated spot). Some spots could correspond to the same DNA molecule. By pooling data from replicates, a more reliable classification of gene expression can be provided and will greatly reduce misclassification rates. The term “replication point” will also be used to describe a replicated spot; and a “*replication set*” describe a set of replicated spots.

**Experiment replication** (repeated hybridization) The same set of spots under the same experimental conditions can be replicated and measured again. In some cases, in order to compensate dye errors, the dyes are interchanged: this last strategy is known as *dye-swap*.

**Empty spots:** Correspond to an spot in which no DNA molecules have been deposited. This type of spot are frequently used to obtain a value for background intensity

**Grid:** The print grid in a microarray image describe the row and column distances, axes and origins to recover the positions of printed spots. Before ratios can be computed and passed on to a data analysis pipeline, appropriate image regions must be mapped to the positions of the print grid or the printed sequences, respectively. This task called ‘gridding’ is usually done using semi-automatic programs

**Slide structure:** Represent a spatial-regions layout of the image –that in this way is organized into ‘partial grids’-, which then enable spatial estimation and distribution of diverse parameters. The application is not able to automatically detect the slide structure (due to different numeration systems used to this end), thus during the “Load Step” this structure is manually specified. The minimum input data apart from the scanned image intensities are the number of rows and columns of each grid, but more side information about the placement of grids can also be used to impose constraints on the allowed segmentation results, yielding more robust processing.



**Slide category:** A set of slides with the same “category mark”

**Functionality** or function or meaning: Describe the PreP meaning of a given label

**Label:** A metadata that describe an specific functionality

**Action,** operation, pre-computing o pre-processing, step. An action have the effect of creating a new layer of pre-processed data.

**Signal:** it refers to the luminosity intensity in the spot.

**Background:** is the unwanted output from the image when operating in the absence of signal - that is in the dark.

**Target and Control:** refers to the two different measurements. Control genes correspond to the reference measurement, and target to the experimental condition.

**Coherent data / no-coherent data,** negative and missing values are un –tractable in the log space. Thus, they are not taken into account in most procedures

**Slide channels.** Red and green channels

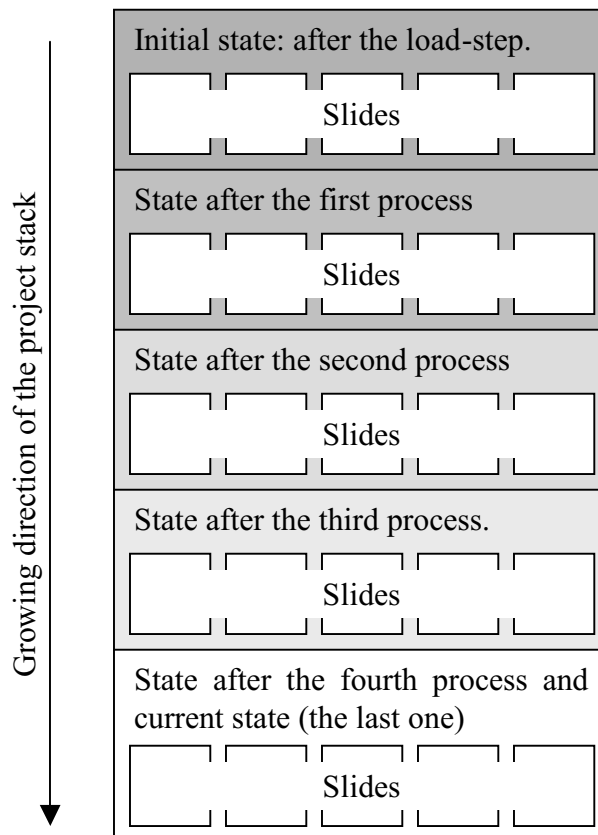
**Internal Note:** *The “Load Step” is going to be simplified in the following way:the program will be prepared to automatically associate columns and functions under the constrain that user follows a given set of labels. In this way, the manual link between columns and functions will remain only as an optional procedure.*



### 3.- Organization of a PreP project

A **PreP** project is a collection of states. Each state is the result of applying a given process over the previous state. The different **PreP** states are stored using a stack, which means, that the last state is the only state that can be removed from the top of the stack: states are pushed into the stack and the last state can be removed (pop) from the stack.

The last state is the *current state*, this is to say, the state over which the procedures are applied (the rest of states conform the “history”). Each state is self-contained, this is to say, they contains all the necessary information to produce a new state (this allows to use a test-error approach to obtain the best results).



**Figure 1. The PreP Project representation.**

Each state is composed of a collection of *slides*. The *slides* represent and contain the information obtained by *scanning* a given DNA chip. The *slide* have an associated name and, when necessary, a set of pre-computed values to be used in a new step. In general, the slide-name resemble the experimental conditions.

A *slide* is a collection of *spots*. Each spot have a set of values that correspond to luminosity intensities, position in the chip, labels, etc. (see Figure 2). The first state is produced by a special step named the “load step”. In this step the slide files are loaded and analysed to



associate the data. Options available for the “load step” are particular to this step (and different for the next “normal” steps).

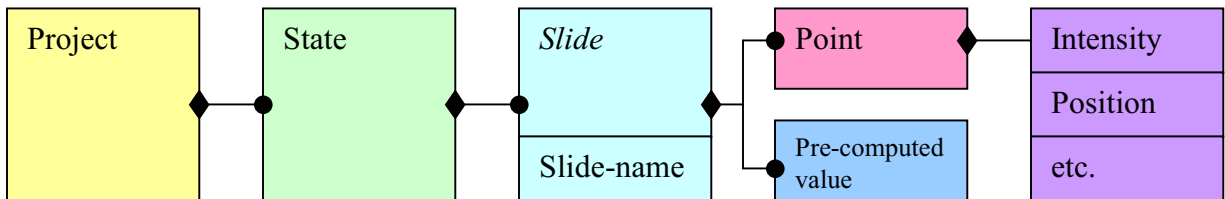


Figure 2. Objects diagram of a **PreP** project. Diamonds represents “is composed of” and circles represents “one or more”.

### 3.1 Working at the Project level

Three different actions are available at project level (and present in the project tools bar):.



Figure 3. The Project Tools bar: New, open and save options

- **New Project:** A new project becomes available as the active project (previous projects, if any, does not disappear).
- **Open a project:** A file dialog window is shown to introduce (or browse) the project filename. The project is open and loaded (previous projects, if any, does not disappear).
- **Save a project:** Store the project on the filename introduced in a file-dialog window. To modify the project filename, specify the new name when save the project.

### 3.2 The Load Step

The available procedures for the load step are : (a) Load a new slide, and (b) End the Load Step. Both actions can be invoked from the tools bar.



Fig. 0, Tools Bar in the Load Step

The objective for the load step is to identify the representative data contained in slide files. These representative data can be: intensities, positions and descriptions (only intensities are mandatory). **PreP** is able to recognize some standard labels and associate these labels with specific functions. However, when a slide file contains other labels different than the **PreP** standard labels, the process of assigning function must be manually completed..



### 3.2.1.- Loading slides and the Screen Layout

Loading the slide is launched by clicking the button and selecting a slide filename using a dialog box. Once the slide has been loaded, labels are displayed and the processes of assigning function to the different labels is started (see *Annexe A, File format for slide* for more information). Next picture represent a typical slide in the “Load Step”, where highlighted fields correspond to: (note: use the pointer over the field for long names)

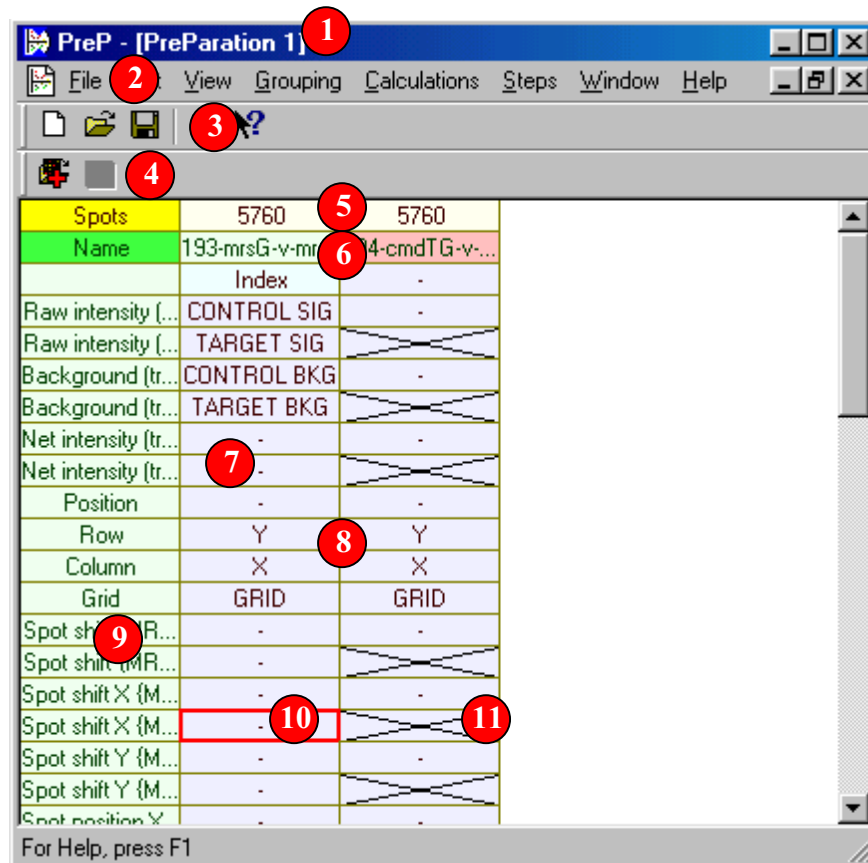


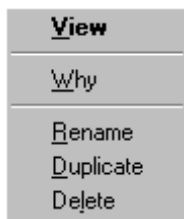
Fig. 0, Screen layout in the Load Step

1. **Project Name:** Is displayed as Window-Title (together with the application name “PreP”) and can be modified by saving the project with a new name.
2. **Menu:** The alternative way to use the tools.
3. **Project Tools bar:** (see below) perform operation at project level, not only over the current state.
4. **Load Step Tool bar:** Contains the options for the load step : add a new slide, and finish the load step (which is turned on when functions has been assigned to labels).
5. **Number of point in the slides:** Correspond to the number of points in each slide. In most cases this value is the same for all the slides, due to they come from the same DNA-array system.



6. **Slides name:** Identify the slide. It can be modified by a slide action. The name displayed in green means that all the mandatory functions (or label meaning) have been assigned and web grouped. The name in red means that the process is not complete. Clicking over the *slide-name* give access to the slide-actions menu (to see the complete slide name put the mouse pointer over the name).
7. **Label without function in the slide:** A label with a hyphen (-) means that there is not function (or meaning) associated to it.
8. **Label meaning (for each slide):** Each entry in the table describes the function or meaning of label (in the left) for this specific *slide*.
9. **Labels:** List the available labels for all the slides. A meaning must be assigned to each label.
10. **Cursor:** Is used to select a label to modify its function. The cursor is controlled with the cursor keys.
11. **Label not available in the slide:** When a given slide does not have a label present in another slide it is represented with a cross, meaning that the label is not available.

### 3.2.2.- Actions over the *slide*



Use the right mouse button to display the menu of “contextual” actions over the slide (see figure on the left) The available actions are:

- **View:** Display *slide* data is the default action (also available by clicking the left button) and switch the visualization mode. Data are displayed as were loaded from the slide file. To switch again to the “load step” visualization mode click again.
- **Why** the *slide* is not complete? : When the slide name is in red means the load step is incomplete. Some mandatory functions have not been assigned or not correctly grouped. This action describe the problem.
- **Rename** the *slide*: A new name can be assigned to the slide. The original slide name is the file name from which it was loaded.
- **Duplicate** the *slide*: In some cases, a slide file contains more than one *slide*. This action allows to assign different functions (or meaning) for different labels in the same file
- **Delete** the *slide*: Remove the slide from the application (and the label-function associations).

### 3.2.3.- Slide data visualization

As mentioned below, it is possible to display the original data by using a slide-action or by directly clicking the left button (default action) over the slide name. Next picture correspond to this visualization. Two fields have been highlighted:

- **Label name:** the column header that describe it.
- **Label Values:** Shows the list of values associated to each label. Each row correspond to one point in the slide thus, the number of rows correspond to the number of points in the slide, except the first rows that contains the label names).



Any mouse-click over the table switch out the visualization mode.

	Raw intensity [...]	Raw intensity [...]	Background (tr...	Background (tr...	Net inte...
1	1929,426316	2176,868421	563,877551	1510,234694	1365,
2	4611,494737	4532,715789	624,591837	1377,234694	398f
3	966,084211	2669,878947	422,571429	2137,857143	543f
4	1406,215789	3755,721053	503,581633	3373,72449	902,f
5	626,426316	4055,089474	551,94898	360,3878	74,4
6	554	4217,3	444,408163	418,438776	109,f
7	6238,668421	9699,168421	580,408163	4199,867347	5658,
8	675,736842	4308,294737	660,204082	4664,153061	15,f
9	710,857895	3585,9	777,55102	4191,22449	
10	583,136842	2514,605263	504,612245	2382,173469	78,5
11	546,473684	2336,684211	776,612245	2674,112245	
12	805,784211	1836,489474	707,204082	2204,938776	98,5
13	17242,08947	15233,87895	432,122449	1739,540816	1680f
14	6214,684211	3301,210526	519,173469	1338,387755	5695,
15	37162,24737	36660,73158	593,102041	2621,77551	3656f
16	1101,005263	3666,878947	712,632653	3462,326531	388,
17	532,678947	3761,110526	610,081633	4011,397959	
18	241,9	4599,126216	502,5	4222,724694	

Fig. 0, slide view. (1) Label names (2) Label values

### 3.2.4.- Available functionality for each label

Before describing the procedure to assign a function (or meaning) for a label in a given slide, let's review the functionalities and the relationship among them. There are mandatory and optional functionalities. Intensities are mandatory, and description and position are optional. However, there are several ways to define these values. Table 1 summarizes the methods. The following are the fields:

**Functionality Type:** Describe one of the available functionality type {intensities, position or description}. Only the first one is mandatory.

- **Specified by...:** Depending on the functionality type, it is specified by its Net values or by the Signal and background values (intensities); or by supplying the (x,y) coordinates and/or the grid position (position feature).
- **Specification method:** Describe the functionality (or meaning).
- **Functionality:** These labels correspond to the description used by the application in the load step.
- **Hot Key:** The key used to assign a given functionality.
- **Description:** a comment about the functionality.



Functionality Type	Specified by...	Specification method	Functionality	Hot Key	Description
Intensities (mandatory)	...one and only one of ...	Net Values	TARGET	T	Net intensity value in the target channel.
			CONTROL	C	Net intensity value in the control channel.
		Signal and background values	TARGET SIG	A	Signal intensity value in the target channel.
			TARGET BKG	R	Background intensity value in the target channel.
			CONTROL SIG	O	Signal intensity value in the control channel.
			CONTROL BKG	N	Background intensity value in the control channel.
Position (optional)	... one or both of ...	Coordinates	X	X	X Coordinate (relative to the grid when it is specified). A natural number (integer $\geq 0$ ).
			Y	Y	Y Coordinate (relative to the grid when it is specified). A natural number (integer $\geq 0$ ).
		Grid	GRID	G	Grid Number. A natural number (integer $\geq 0$ ).
Description (optional)		Description	The description name	ENTER	Take the label values as a description.

**Table 1, Available functionalities for each label.**

It is noteworthy observe that to assign functionality it is necessary to introduce a description name (label value). This association is maintained along the process. During the “save” step the user can decide what descriptions will be included in the output.

### 3.2.5.- Assigning function to labels

There are three methods to assign function to a given cell.

1. **Assignment Keys:** The hot keys described in the table are the fastest method to assign function to the current active label.
2. **Cell menu:** Press the right button on a given cell and a pop-up menu becomes available. Use the cursor or the mouse to select the function to be assigned.
3. **Label menu:** By clicking on the label name the function is assigned to all the slides.

It is also possible do not assign functionality to certain cells (*unused*). Next pictures shown the cell and label menus..

Go to	
Unused	Del
Tag...	Enter
Target	T
Control	C
Target Signal	A
Control Signal	O
Target Background	R
Control Background	N
X	X
Y	Y
Grid	G

**Fig. 0, Cell Menu.**

Unused
Tag...
Target
Control
Target Signal
Control Signal
Target Background
Control Background
X
Y
Grid

**Fig. 0, Label Menu.**



### 3.2.6.- Ending the load Step: the *slide* structure

When all the *slides* have been loaded and the mandatory functionalities assigned to each slide, the load step can be finished. At this point, if the position functionalities have been defined, the application need additional information about the structure of the *slide*.

This slide structure can not be directly deduced by the application due to the positional indices can be numbered from 0 or from 1 and the *grids* can also be numbered in several ways. The following dialog-box is used to introduce the *slide* structure:

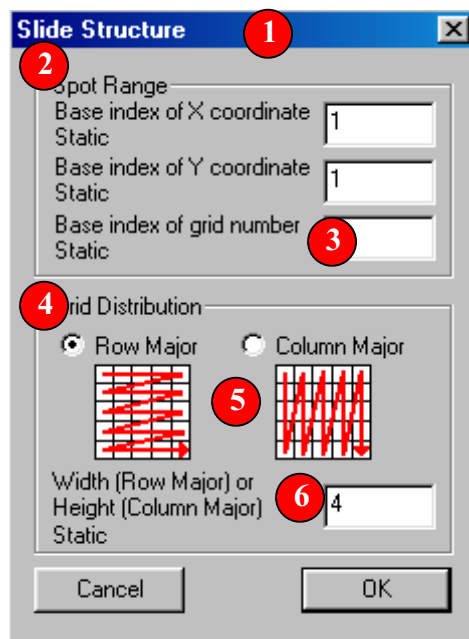


Fig. 0, Dialog box to describe the slide structure.

1. **Title:** The name of the *slide* to be structured.
2. **Coordinates range:** Need to be user defined since the origin of coordinates can be 0 or 1 in dependence of the program used to analyze the scan-image.
3. **Base values:** There is a “base-value” for each coordinate and for the numbering of grids. The default value is the minimum value in each range (although this value could be different when empty rows or columns are present).
4. **Grids distribution:** Information needed to deduce the spatial grid distribution.
5. **Grids order:** Grids can be organized from left to right and top to down (*row major*) or from top to down and left to right (*column major*).
6. **Distribution Size:** Is the maximum number of possible alignments, and it correspond with the width in *row major* or the height in *column major*.

Next examples will illustrate the different options available in the dialog box. Lets be the following data set (only the position label are shown):



Gen	X	Y	Grid
A	1	1	1
B	2	1	1
C	1	2	1
D	2	2	1
E	1	1	2
F	2	1	2
G	1	2	2
H	2	2	2
I	1	1	3
J	2	1	3

Gen	X	Y	Grid
K	1	2	3
L	2	2	3
M	1	1	4
N	2	1	4
O	1	2	4
P	2	2	4
Q	1	1	5
R	2	1	5
S	1	2	5
T	2	2	5

Because the minimum value is 1, we choose 1 as Base-value and for the grid distribution, *row major* with width 2. Thus, the gene position will be:

A	B	E	F
C	D	G	H
I	J	M	N
K	L	O	P
Q	R		
S	T		

If we change the width to 3, the organization will be.

A	B	E	F	I	J
C	D	G	H	K	L
M	N	Q	R		
O	P	S	T		

Using *column major* with high=2, the data set is distributed as follow:

A	B	I	J	Q	R
C	D	K	L	S	T
E	F	M	N		
G	H	O	P		

However, if the base of coordinate X were 0, the first values will be taken as empty (due to the data start in 1), resulting, with *row major* of width 2:



	A	B		E	F
	C	D		G	H

	I	J		M	N
	K	L		O	P

	Q	R			
	S	T			

Finally, using base 0 for the *grid*, then, the first grid will be empty (the number 0 without data):

		A	B
		C	D

E	F	I	J
G	H	K	L

M	N	Q	R
O	P	S	T



## 4. - Slides Visualization

This section is devoted to the different slide visualization procedures, and how they can be grouped, selected, etc. Some of these action are necessary to allow an operation to be applied over a given state. Since the state organization is a stack, the last state is the only available –at a given time- to deal with. In fact, all we are going to describe is only applied over the last state in the stack: the current state.

### 4.1.- Screen Layout

Next picture shows the screen layout and the main elements in the state view.

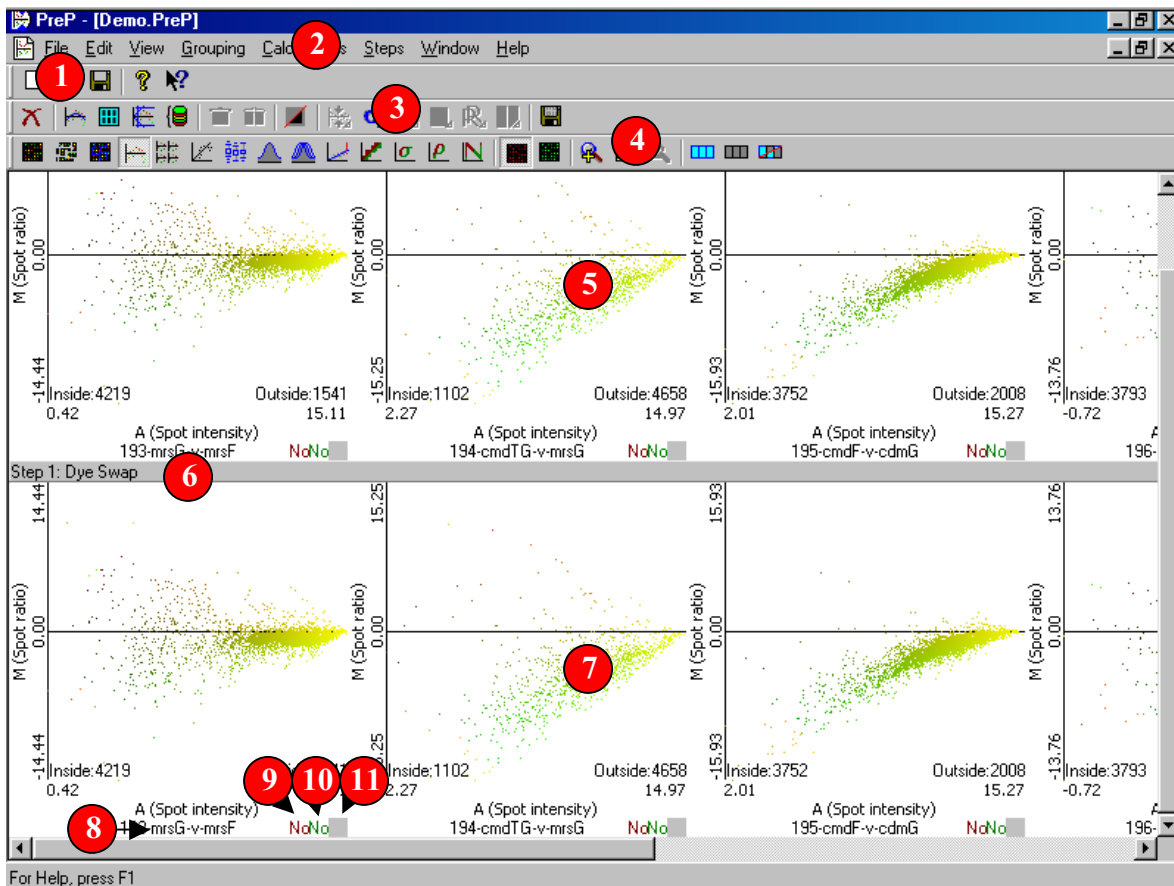


Fig. 0, Screen layout and visualization steps.

The view is composed by a visualization matrix. Each row corresponding to one state, being the last the current state (from top to down), and the only one over which the procedures are applied. At the same time, each row is composed by one view for each of the different *slides* that conform the state. Some procedures can change the number of *slides* from one state to the next. The different elements shown in the picture are:



1. **Project tool bar:** (see below), to create a new project, open an existing project or save the current project.
2. **Menu:** to select procedures, that are also available in the tools bar (see above).
3. **Tools bar to operate over the current state:** Contains actions to be applied over the current state. The following main task can be used: delete state, preparation of one operation, steps (operation over the state) and saving data.
4. **Tools bar for visualization and selection:** Contains actions related to visualization and selection. Is composed by visualization type, visualization channel, zooms and selection.
5. **Slides visualization in the previous states:** for visualization of states different than the active (last row). Any change on the visualization mode not only affect the current state, but also the previous.
6. **Step number and operation name:** is the title assigned to each state and is formed by a correlative number (state 0 correspond to the load step) and the action producing that state.
7. **Viewing slides of the current state:** Last row of views correspond to the current state (to whom will affect the actions).
8. **Slide name:** Is displayed in the bottom part of the view. This name is carried along the different states.
9. **Replicated Points:** For the *intraslide* replication operations it is necessary to group points forming replication sets. Each replication set correspond to a measure that have been taken several times (a replicated point). This number is the number of replication sets in the slide.
10. **Slide mark:** *slides* can be marked. These marks are used in some actions, such as in the *dye-swap*, to represent dyes has been interchanged
11. **Slide group:** in some actions such as the *dye-swap* it is necessary to group the slides.

## 4.2.- Viewing types

The different type of visualization has been grouped in the viewing toolbar:



Fig. 0, Viewing type options in the tools bar.

The icons, from left to right are:

- **Slide view:** It is a synthetic image, built up from the available data. When data are not available, the spot is drawn with a grey mark
- **Coherent slide view:** It is a synthetic image, built up from the coherent data. This means that negative intensities and not available data are not shown.
- **Slide view with quality:** The quality value (supplied by replication algorithms) is shown in blue in this view.

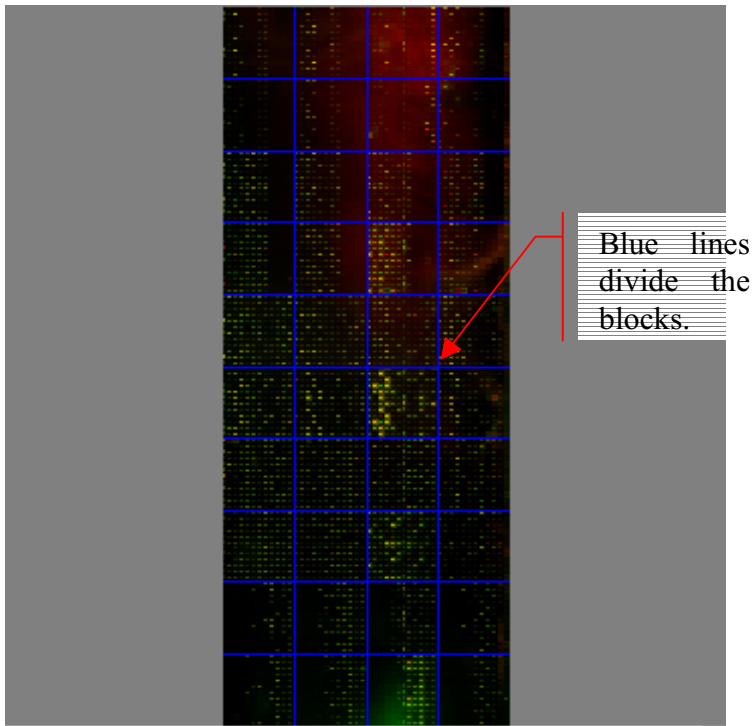


- **AM Graph:** display the points by intensity A (axis x) and color M (axis y). This graph allows to establish the dependency of the color with respect the intensity, which is in fact, one of the most typical problems in normalization applications.
- **AM Graph by blocks:** Display the AM graph for each of the blocks in which the slide is divided (by default using the grid). This graph allows identify variations between the different zones of the slide.
- **RG Graph:** Display the point from the red channel R (axis x) and the green channel G (axis y). The election for the channels is always red for the target and green for control.
- **Box Graph:** A box graph displays where half of the data are concentrated and makes an estimation of the possible variation, showing the data out of range independently. This graph is performed for each block in the slide, and allows visualise the differences in range for each block.
- **Values density:** Display the probability density curve of the colour value distribution. Ideally this values should be centred around zero in the normal form.
- **By block Values density:** the same as the previous graph, for each block in the slide. This allows to study the effect of the position in the slide.
- **Intensity-Intensity Graph:** Compares the intensity between two slides, giving an approximate idea about the quality of the double-scan and data dependencies between them.
- **Scatter plot of replicates:** When several points has been grouped it is possible its visualization with respect to the average of the whole set. Ideally they should be distributed around the bisect.
- **Standard deviation of the replicates:** shows the SD for each replication set with respect to the average value. Ideally they should show a linear relationship.
- **Replicates correlation:** Display the correlation between the two channels (target and control) for each replication group. Ideally it should be 1.
- **Normality of replicates:** Compute the distribution function for each of the replicate sets and drawn it with respect to the normal distribution. Ideally it should be a identity function, this is to say, it should move along the bisect.

The default viewing is the AM graph (highlighted in the previous picture).

#### 4.2.1.- Slide view

Is a synthetic reconstruction of the slide image, from the available data. If data are not available for a given spot a grey mark is set. The block structure of the slide is also displayed. The layout of this view is shown in the next picture:



Blue lines are used to divide the slide in blocks (initially they correspond to the grid specified during the “load step”). Detailed information for each point is supplied in a box, when the cursor is set over the point (see next picture).

Fig. 0, Slide View

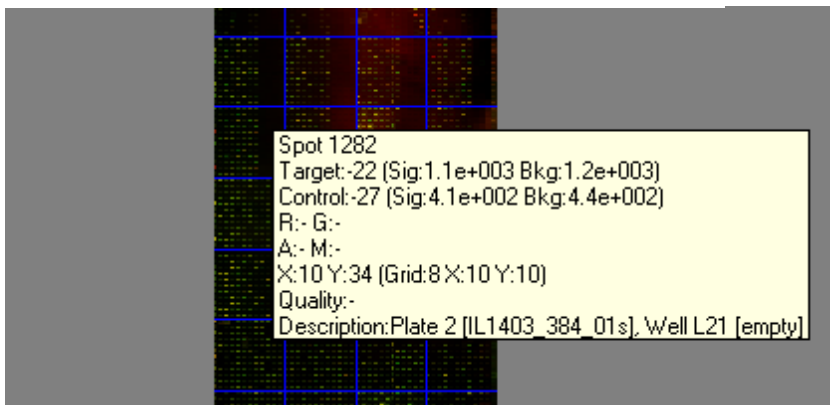


Fig. 0, Associated information for each spot.

The displayed fields means:

- **Spot:** The spot number.
- **Target:** Intensity in the target channel (red), first the net value, then (in brackets) the signal and the background.
- **Control:** Intensity in the control channel (green), first the net value, then (in brackets) the signal and the background..
- **R:** Log of the net intensity in the target channel (when positive)
- **G:** Log of the net intensity in the control channel (when positive).
- **A:** logarithm of geometrical average intensity [ $A=1/2(R+G)$ ]



- **M**: logarithm of the ratio intensity  $M=(R-G)$ .
- **X**: x coordinate in the slide.
- **Y**: y coordinate in the *slide*.
- **Grid**: Grid number and position inside the grid.
- **Quality**: spot quality
- **Tags**: Name of the labels associated to this spot and its values

When a point belongs to a replication set the additional information corresponding to the replication set is also displayed:

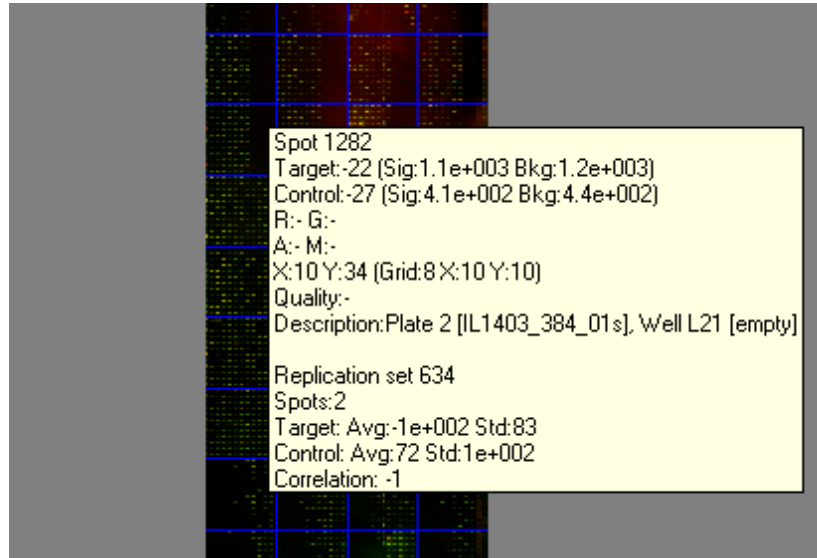


Fig. 0, Associated information for a replication set.

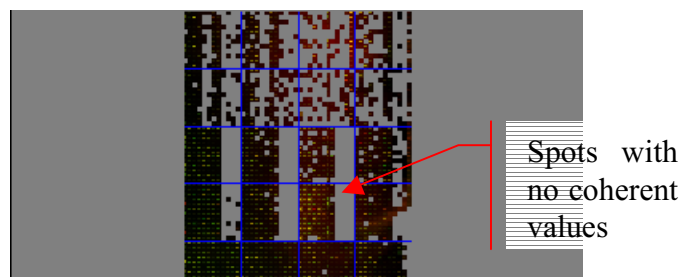
These values are:

- **Replication set**: The number of the replication set.
- **Spots**: The number of points in the replication set.
- **Target**: Average and standard deviation of the replication set for the target intensity
- **Control**: Average and standard deviation of the replication set for the control intensity
- **Correlation**: between the target and control intensity values in the replication set.

#### 4.2.2.- Coherent *slide* view

Similar to the previous view but only coherent data are displayed. This means that negative intensities and not available data are not shown.

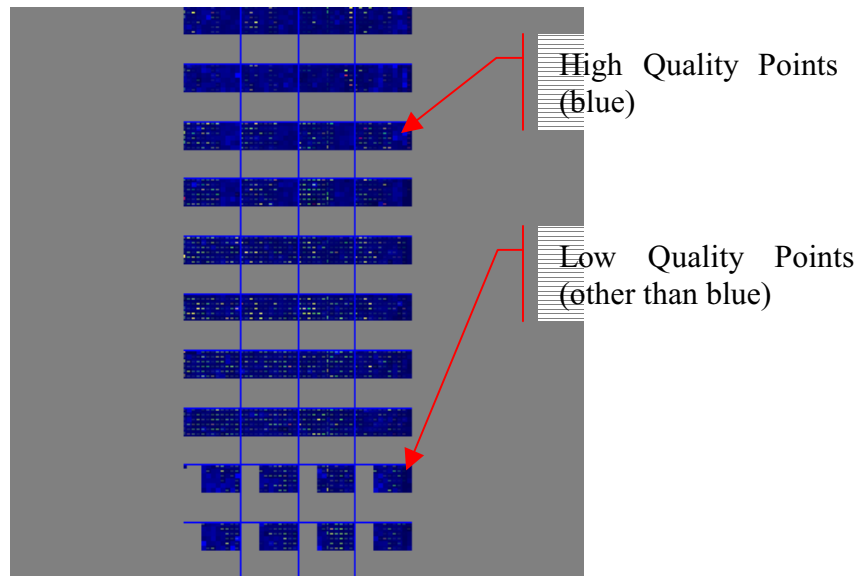
Detailed information for each point is also available through the mouse pointer.





#### 4.2.3.- Slide view with Quality

The replication algorithms include a quality measure about results. This value is shown in blue in the “view of quality” (it is also displayed for the spot in which the pointer is).



**Fig. 0, Slide view with quality information**



### 4.2.4.- AM Graph

Display the points in intensity A (axis x) and colour M (axis y) coordinates. This graph allows to establish the dependency of the colour with respect the intensity, which is in fact, one of the most typical problems solved by normalization.

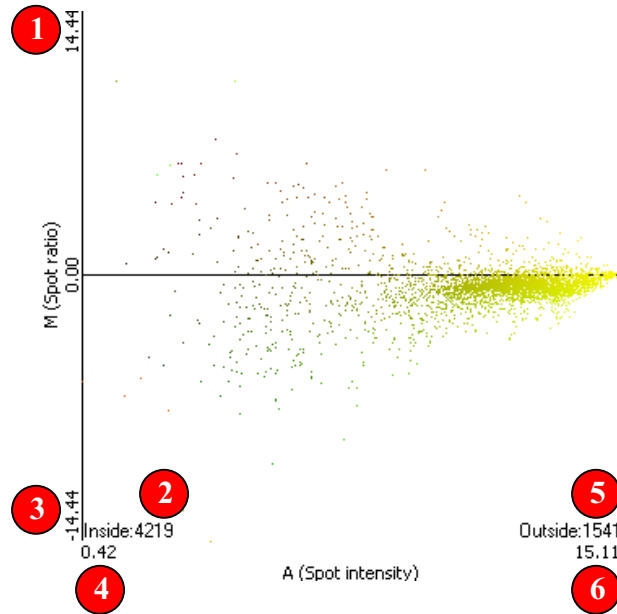


Fig. 0, AM Graph

1. **Maximum value of M:** where M is the log of the channel intensity ratio (log (R/G))
2. **Number of points in the graph:** filtered from the initial set or in the successive steps.
3. **Minimum value of M:**
4. **Minimum value of A:** where A is the log of the product intensity in both channels.
5. **Number of points not represented:** Those spots whit negative intensities or with undefined log are not shown.
6. **Maximum Value of A:** from the spots represent in the graph.

Different views are also available for pre-computed *ratio* can be directly observed from the data.

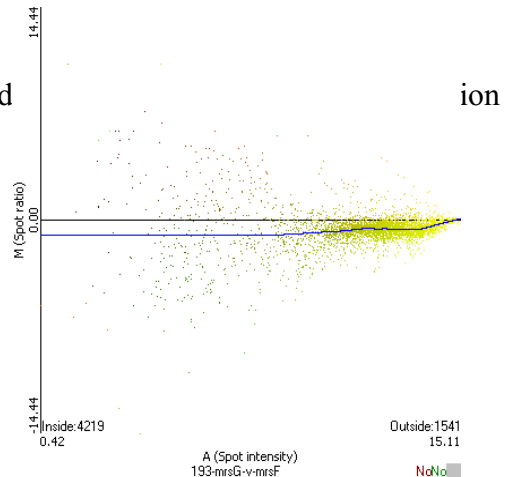


Fig. 0, AM Graph with adjust



There is also available a pre-view of the thresholds used in the filtering steps, as shown in the next picture

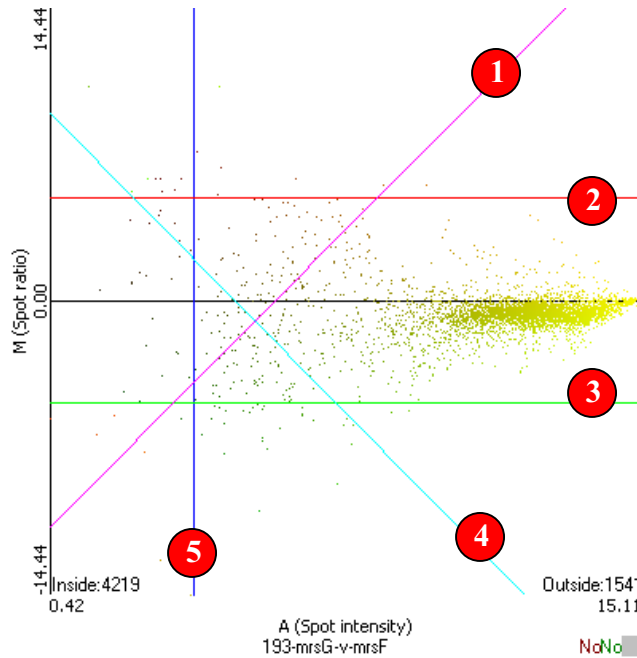
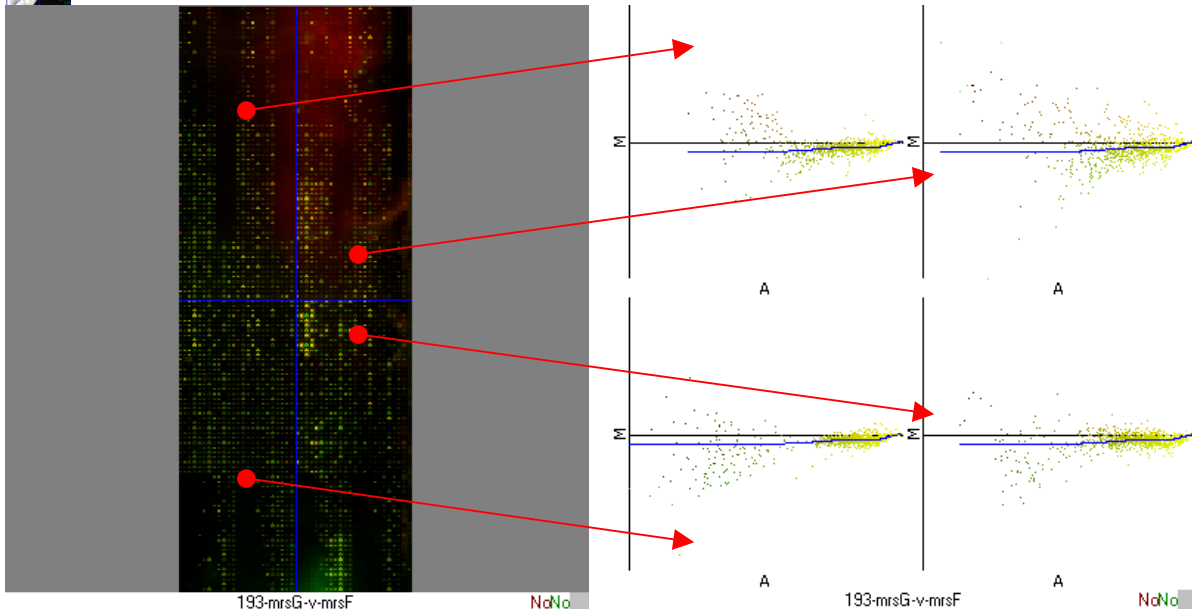


Fig. 0, AM Graph and Thresholds for filtering.

1. **Minimum value threshold in the target channel:** Those points with target value less than the threshold will be removed in the filtering operation.
2. **Maximum ratio threshold:** Points with ratio value greater than threshold will be deleted in the filtering operation.
3. **Minimum ratio threshold:** Points with ratio value less than threshold will be deleted in the filtering operation.
4. **Minimum value threshold in the control channel:** Those points with target value less than the threshold will be removed in the filtering operation.
5. **Minimum intensity value threshold:** Points with value less than threshold will be removed deleted in the filtering operation.

#### 4.2.5.- AM Graph by blocks

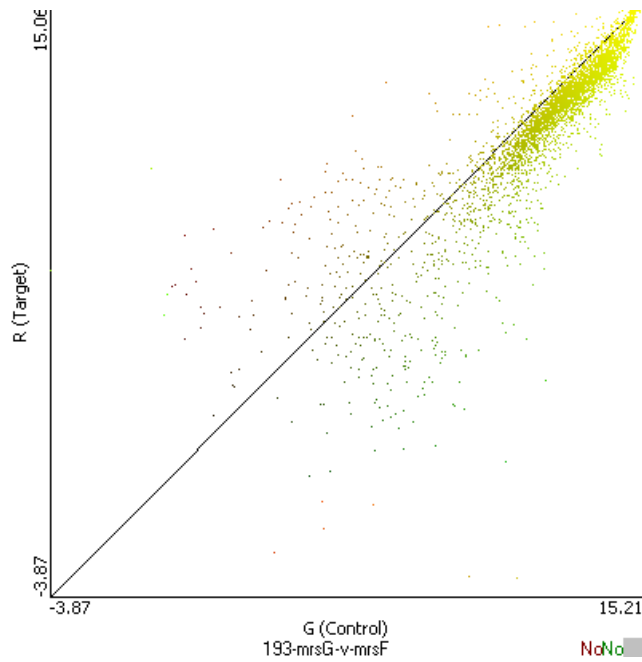
Display the AM graph for each of the blocks in which the slide is divided (by default using the grid). This graph allows identify variations between the different zones of the slide.



**Fig. 0, Slide view and related AM graph by blocks. The relationship between both representations also shown.**

#### 4.2.6.- RG Graph

Displays the points from the red channel R (axis x) and the green channel G (axis y). The election for the channels is always red for the target and green for control. This graph is a rotated version of the AM graph, highlighting the composition of each spot in both channels



**Fig. 0, RG graph.**

#### 4.2.7.- Box Graph



A box graph displays where are concentrated half of the data and makes an estimation of its possible variation, showing the data out of range independently. This graph is performed for each block in the slide which allows visualise the differences in range for each zone of the slide.

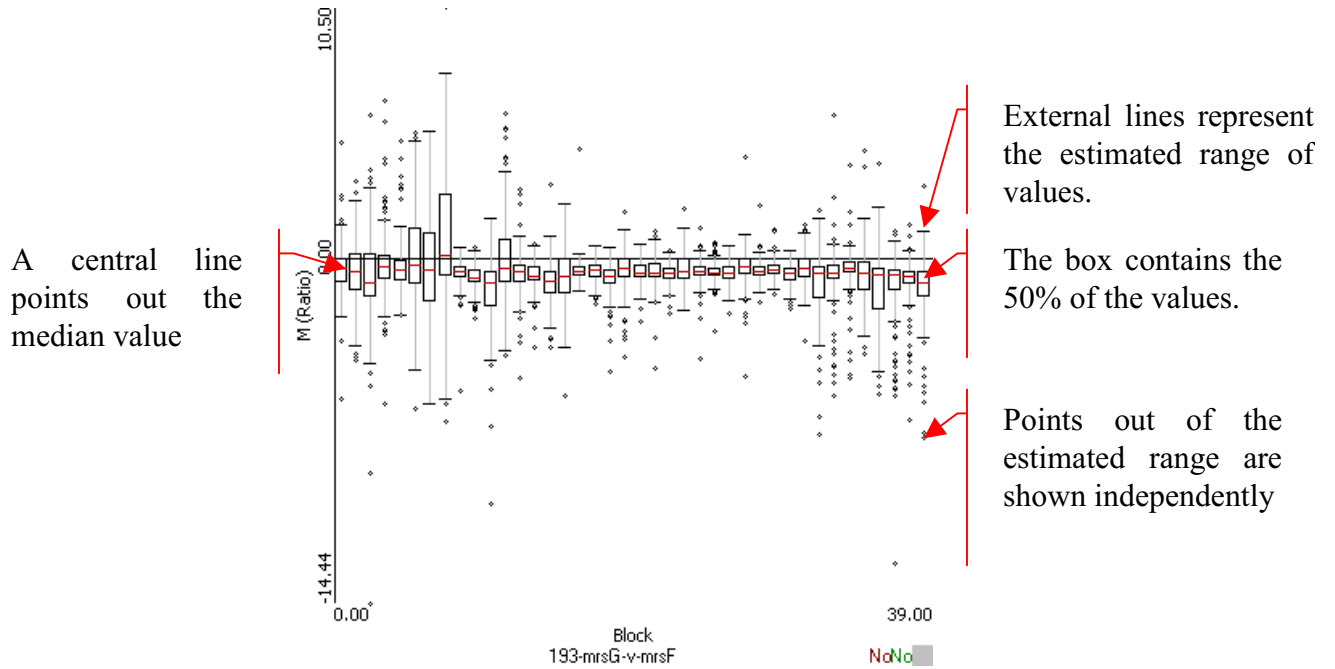
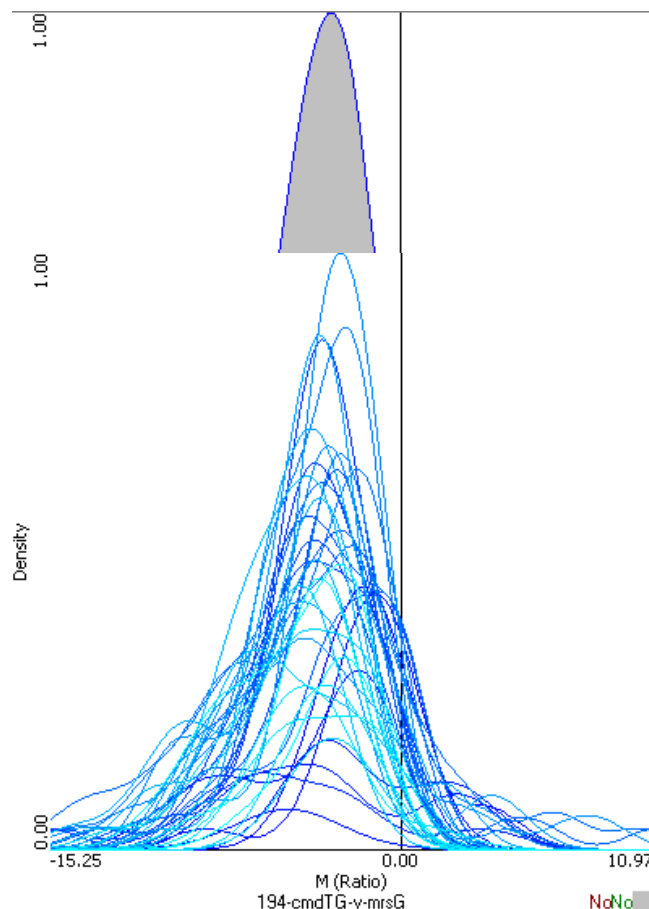


Fig. 0, Box graph.

#### 4.2.8.- Values Density

Displays the probability density curve of the colour value distribution. Ideally this values should be centred around zero in the normal form. Density is a relative measure in such a way that the maximum value is always 1.



#### 4.2.9.- By block values density

Fig. 0, Ratio density values distribution by blocks



Similar to the “density values” graph, but computed for each block in the slide. This allows to study the effect of the position in the slide.

#### 4.2.10.- Intensity-Intensity Graph half of the data

Compares the intensity between two slides, giving an approximate idea about the quality of the double-scan and data dependencies between them. In order to use this graph, two slides must be grouped, and put a mark in one of them. The utility of this representation is for scans of different intensity of the same slide, in such a way that it is possible to observe the influence of the scanner and the non-linear behaviour of the photo-detectors.

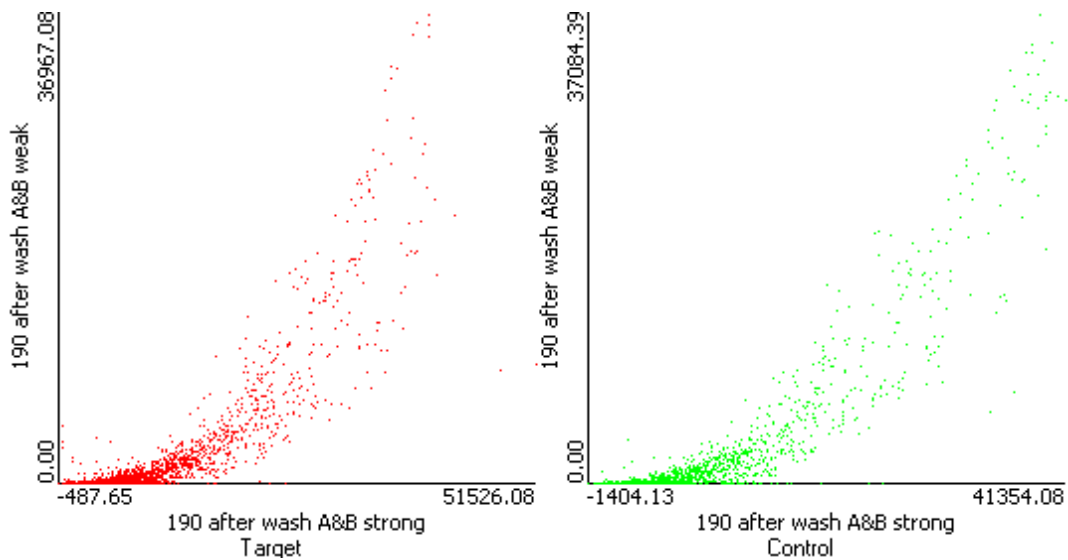


Fig. 0, Intensity-Intensity Graph.

#### 4.2.11.- Scatter Plot of replicates



When several points has been grouped it is possible its visualization with respect to the average of the whole set. Ideally they should be distributed around the bisect.

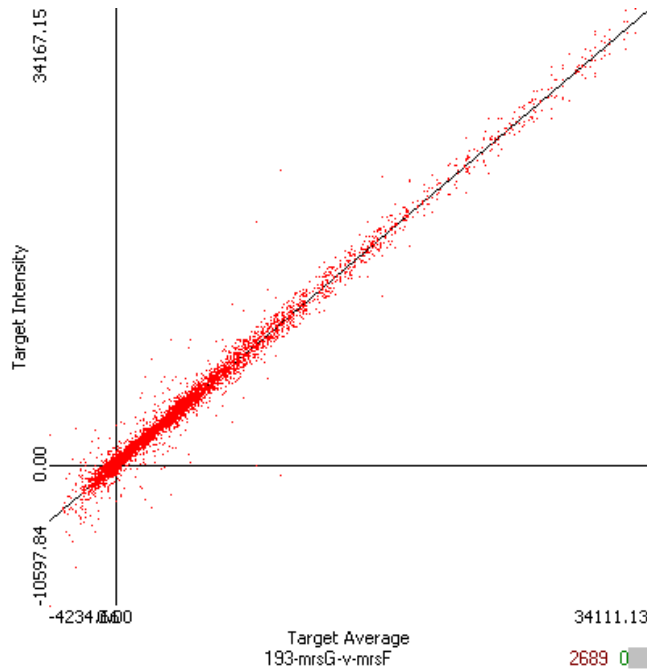


Fig. 0, Scattering of replicates

#### 4.2.12.- Standard deviation of replicates

Displays the SD for each set of replicates with respect to the average. Ideally they should show a linear relationship

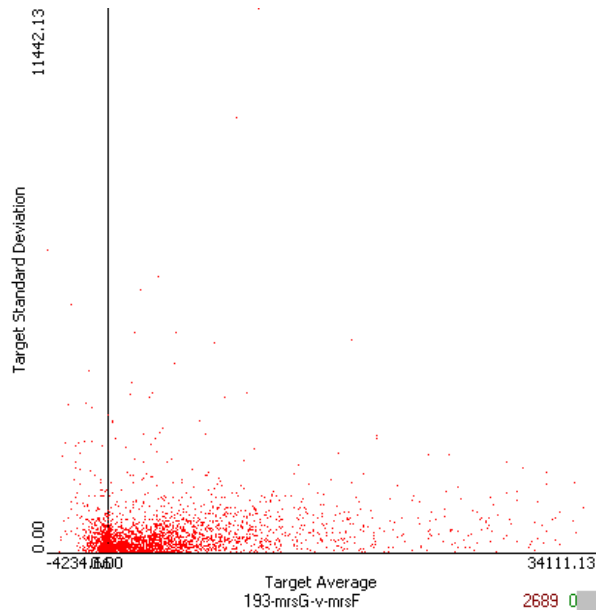


Fig. 0, Standard deviation of replicates against the average value.

#### 4.2.13.- Replicates Correlation

Displays the correlation between the two channels (target and control) for each replication group. Ideally it should be 1.

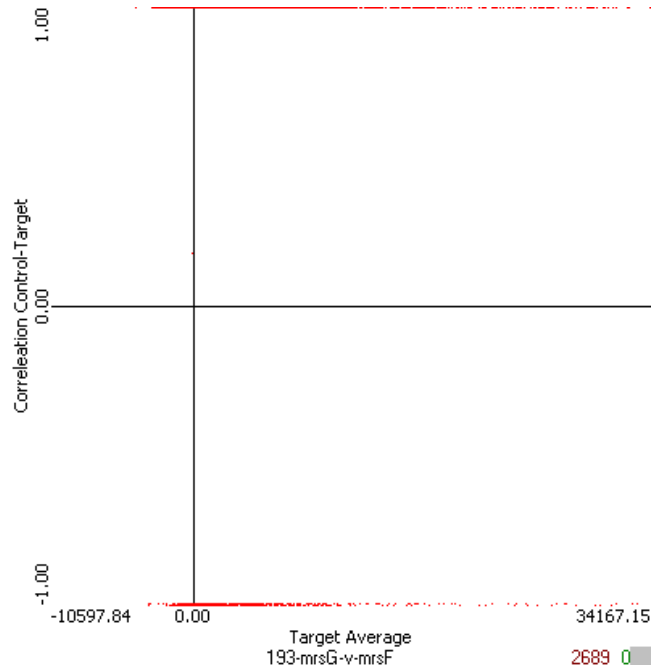


Fig. 0, Correlation against the replicated average

#### 4.2.14.- Normality of Replicates

Compute the distribution function for each of the replicate sets and drawn it with respect to the normal distribution. Ideally it should be a identity function, this is to say, it should move along the bisect.

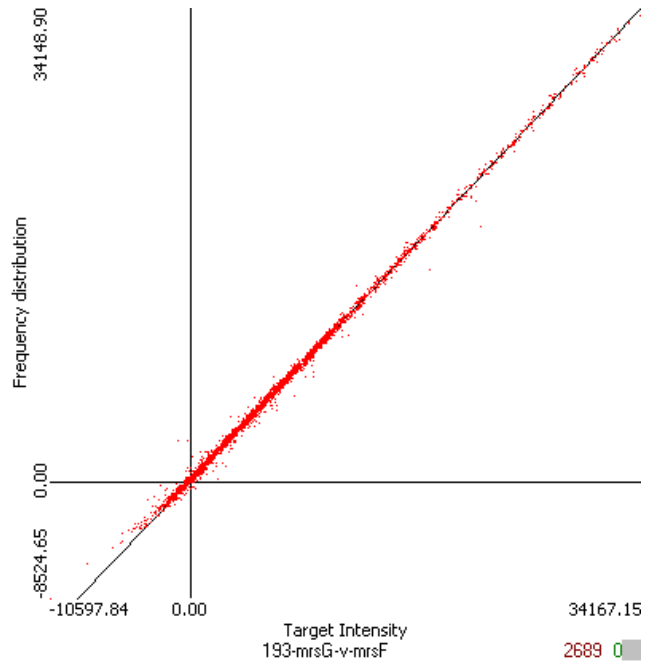


Fig. 0, Normalized distribution of replicates.

#### 4.3.- Visualization Controls

##### 4.3.1.- Slide Channels

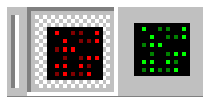


Fig. 0, Channel Selection



There are two intensity values associated to each point in the slide, and they correspond to measures in the two channels of the scanning device. Some of the previous graphs can represent data from the target channel or from the control channel. To select one or other channel, action-buttons are available in the right hand side of the visualization tools bar.

In all the previous pictures, the red channel (target) have been used (when the control channel is selected, points will be shown in green).

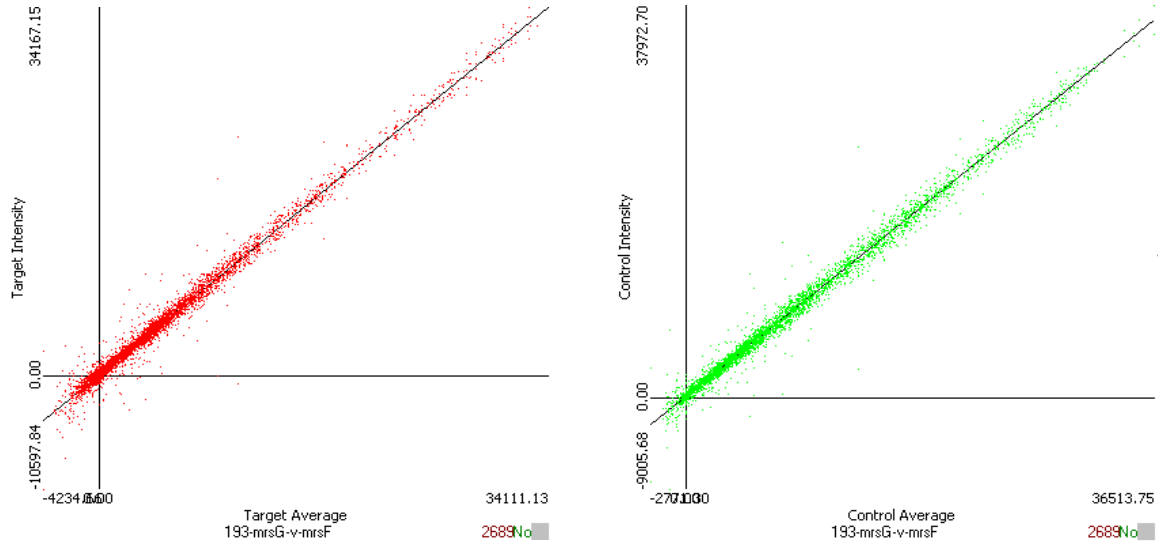


Fig. 0, Replicates scatter plot for the two channels.

#### 4.3.2.- Zoom



Fig. 0, Zoom Actions

Three actions are available in the tools bar:

From left to right,

- **Zoom In** : Increase the views size
- **Full Screen Zoom**: Adjust the view size to the screen size.
- **Zoom out**: Reduce the view size.

Both, *zoom-in* and *zoom-out* are bounded in the scale, and when active, the *zoom* scale have effect over all views (not only over the last state views).



## 5.- Selection, groups and Slide marks

### 5.1.- Slides selection

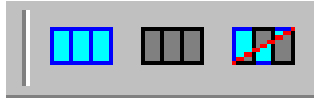


Fig. 0, Slide selection actions toolbar

Any slide of the last state can be selected by clicking over its view. In some visualization modes, where the graph does not represent one slide (i.e. intensity-intensity graph) it is not possible to use this selection method.

The selection action have been (grouped in visualization tools bar) are, from left to right:

- **Select all the slides:** overcome the current selection and select all *slides*.
- **Un-select all the slides:** overcome the current selection and set off all *slides*.
- **Invert current selection:** Change the select mode of the slides (turning on the off slides and vice-versa).

### 5.2.- Grouping slides

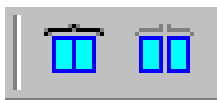


Fig. 0, Grouping actions toolbar.

Slides marked as selected can be grouped. There is not limit for the number of different groups that can be formed, but one slide can only belong to one group. Actions to perform the grouping are in the tools bar, from left to right:

- **Group the selected slides:** Selected slides will be grouped. If a given slide currently belongs to other group it will be moved from it.
- **Un-group previously selected slides:** Selected *slides* will be moved out from those groups they currently belong.

Grouping slides is a mandatory action for specific views, operations and steps.

### 5.3.- Setting Marks to slides

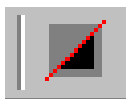


Fig. 0, Marking Action

*Slides* can be marked as relevant slides or as belonging to a particular category. This action is necessary for some views, steps or pre-computing (i.e. intensity-intensity). It is available from the tools bar (Step tools-bar):



This action change the mark-state of the selected *slide*. All the marked *slides* in this selection will be switched to un-marked and vice-versa. However, an easy and faster way is provided through the left mouse button over the marked views.

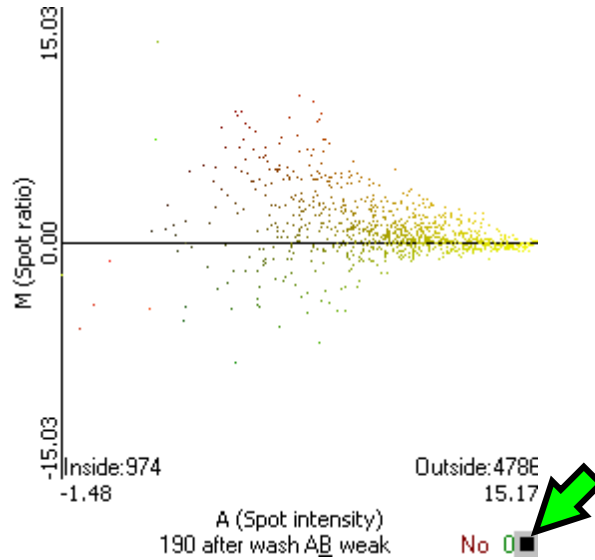


Fig. 0, Mark position in the slide view.



## 6.- Pre-Processing

Pre-processing are procedures performed over the last (current) state. They are necessary as a previous condition to proceed with an operation that produce a new state (step operation). Pre-processing procedures are available from the views to allow the user a visual inspection of data before proceed with pre-processing. Pre-processing action are also available in the Step-tools bar:



Fig. 0, Preprocessing actions in the tools bar.

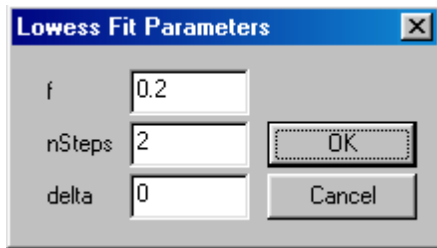
From left to right:

- **Lowess Adjust:** Fits a lowess curve to estimate –in each block- the deviation of the ratio with respect to zero. This is a necessary action previous to the ratio adjust step.
- **Block size selection:** Divide a *slide* in blocks, that are separately processed in some procedures (to account for spatial effects).
- **Set thresholds:** A necessary action for the filtering step
- **Grouping replicated spots:** A necessary previous action to solving replicates. Group together those spots with the same label value, under the hypotheses they belong to the same DNA molecule.
- **Double scan regression:** A necessary previous action to solving the double scan approach. The relation between two slides is adjusted dos *slides* under the hypotheses they correspond to two different readings of the same chip [*This procedure is under patent process, thus it is not available in this application*]

These procedures are applied over the selected *slides*, except in the double *scan* procedure, that requires grouping the *slides* and mark one of them as low sensitivity *scan*.

### 6.1.- Lowess Adjust

This procedure adjust the data to a lowess curve. Since data are a “cloud” of points the adjust will never be perfect. *Lowess* performs a robust adjust, this is to say, closer points are considered (and outliers are discarded). This is quite adequate for those experimental data in which only a small number of data shows different values (which is the case of most gene expression data, in which we expect only a minor number of genes are expected to have differential expression). In this way, the curve adjust to zero those genes with similar expression values.



Lowess procedure needs three parameters:

Fig. 0, Lowess Adjust parameters

- **F Parameter:** Is the fraction of relevant points. As was previously mentioned, lowess adjust assign a greater weight to close points (as a function of the pairwise distance). This parameter set the number of points to be considered as close. For high values of F, the curve is smoother. Usually this value is set in the range [0.2 , 0.5].
- **NSteps parameter:** Number of algorithm iterations. In each step, the curve is better adjusted and smoother. As the algorithm proceed it will trend towards the closer points. Usually this value is set in the range [2 , 10].
- **Delta parameter:** When the number of steps is high (i.e. greater than 100) this parameter speed-up calculations. Otherwise, it should be zero.

Next picture shows the effect of F parameter.

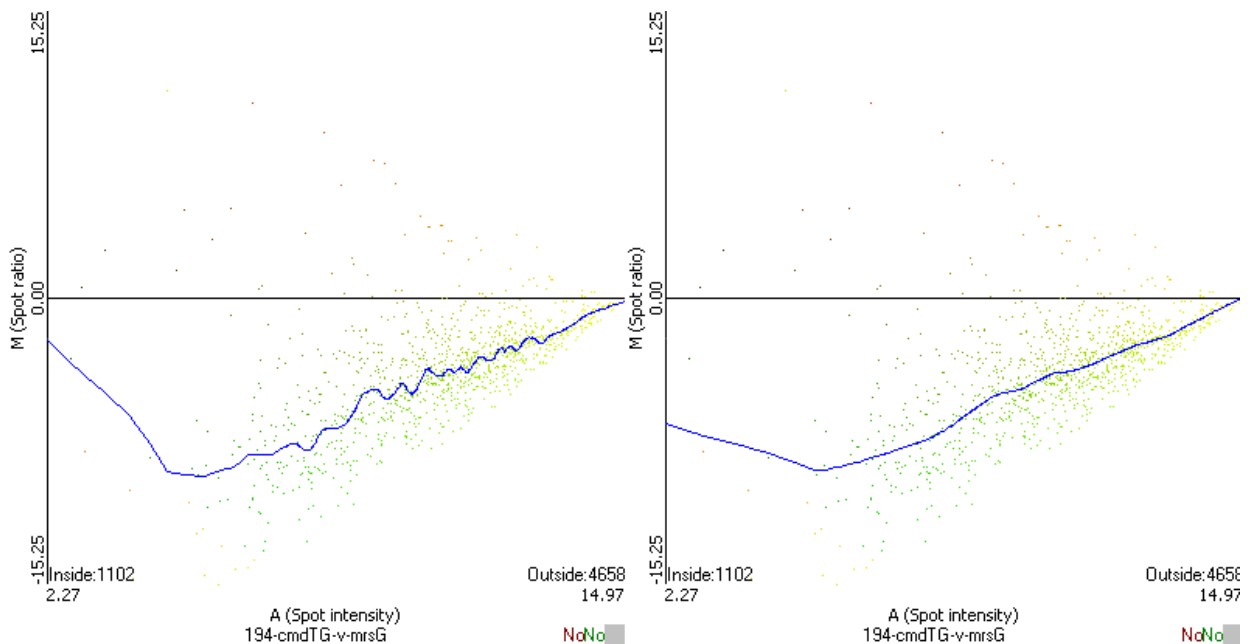


Fig. 0, The same data set with lowess adjust. On the left,  $f=0.05$ , and on the right  $f=0.2$ .

Lowess adjust is performed in each block of the *slide* in order to account for the possible spatial effect.



## 6.2.- Selection of the block size

Splitting slides in blocks is always performed in blocks of the same size. It is also possible to use blocks of the *grids* size, enabling to establish a relationship between blocks and the number of points (spots) used during the printing process (*print-tip group*). These options are available in the Dialog Box for Block Selection:

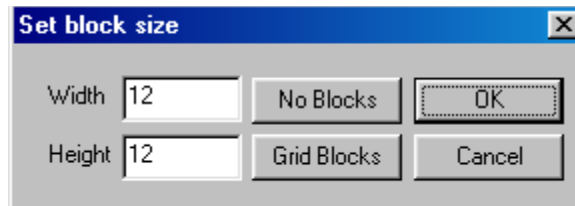


Fig. 0, Selection of the block size.

The blocks partition can be visualised in the *slide* view:

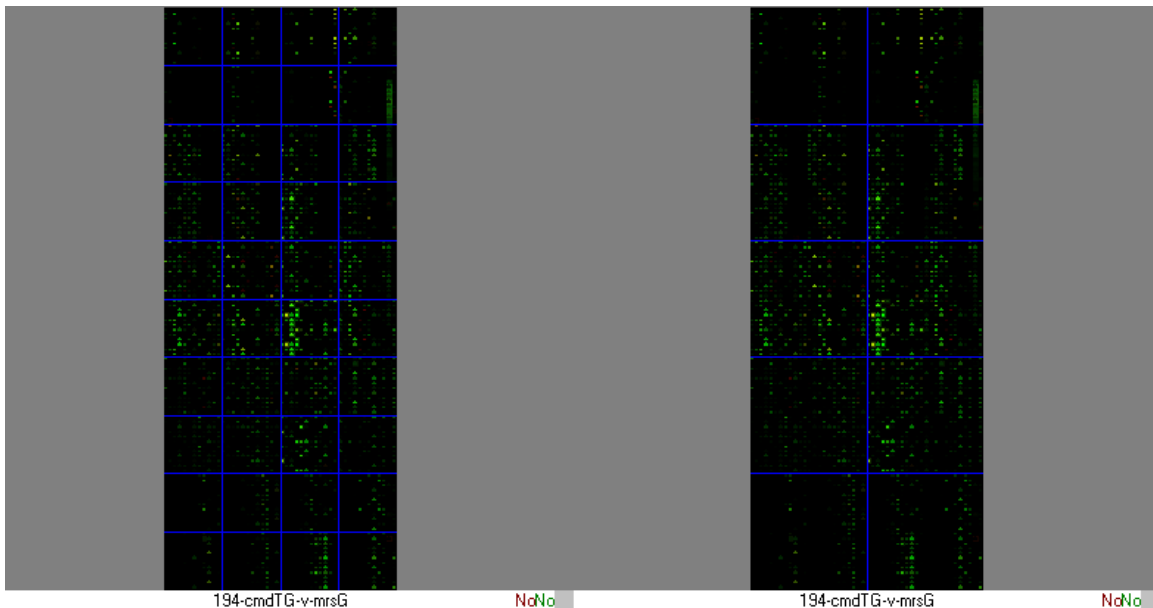
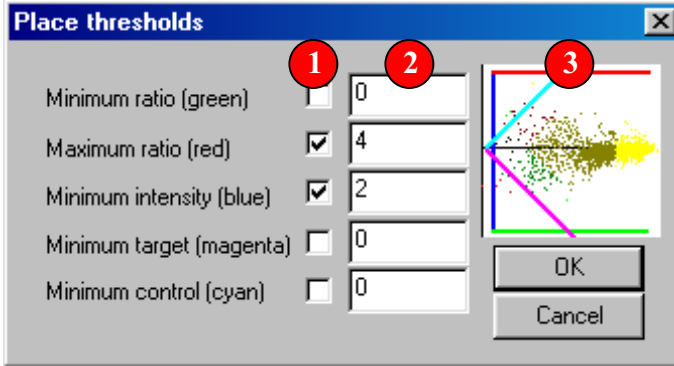


Fig. 0, Different block-partitioning for the same slide. On the left using 12x12 sized blocks and on the right using blocks of 24x24

## 6.3.- Set Thresholds

For the filtering step a threshold is needed. Thresholds are set (in this action) using a dialog box with the following components:

1. **Activate thresholds:** To use a given threshold it must be active.
2. **Threshold value:** A numeric real value.
3. **Threshold Guide:** Identify the thresholds by colour. The position in which the threshold is displayed is only a “guide”, not the real value.



The threshold values are set in the axis of the AM graph.

Fig. 0, Dialog Box for Thresholds.

### 6.4.- Grouping replicate points

Additional to the slides clustering, it is also possible group together some spots inside the slide. Due to the high number of points in the slide, it is not effective to proceed by selection, thus, the grouping is performed using the "labels" assigned during the load step. A dialog box is used to request the "Label" that will be used as identifier of the replicated set. Those points having this label value will become part on that group. Once grouped, the number of groups is displayed in the right bottom of the view.

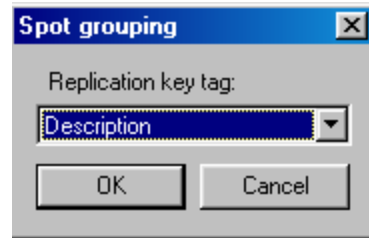
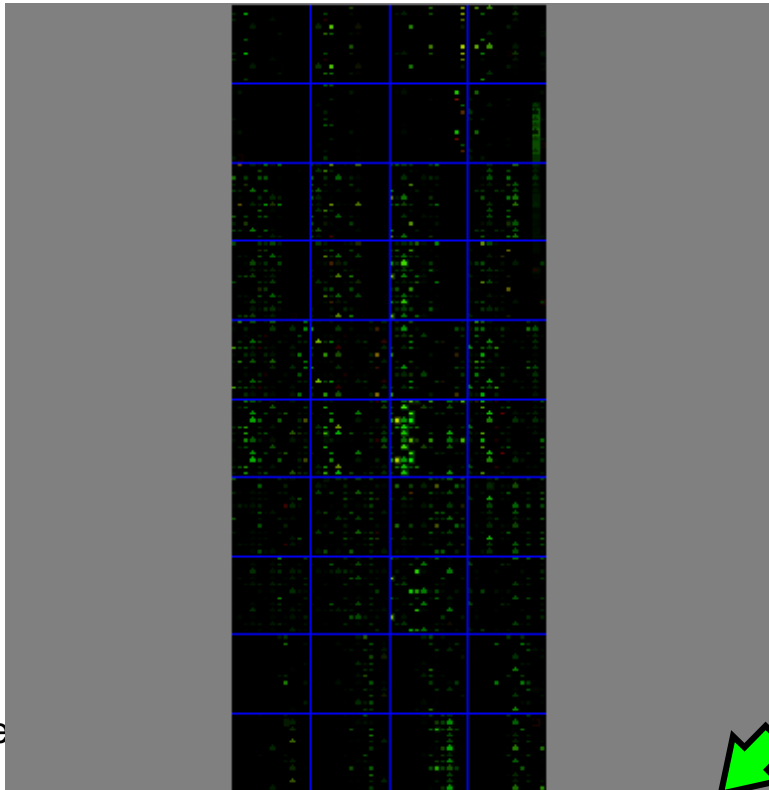


Fig. 0, Dialog Box for grouping replicates.



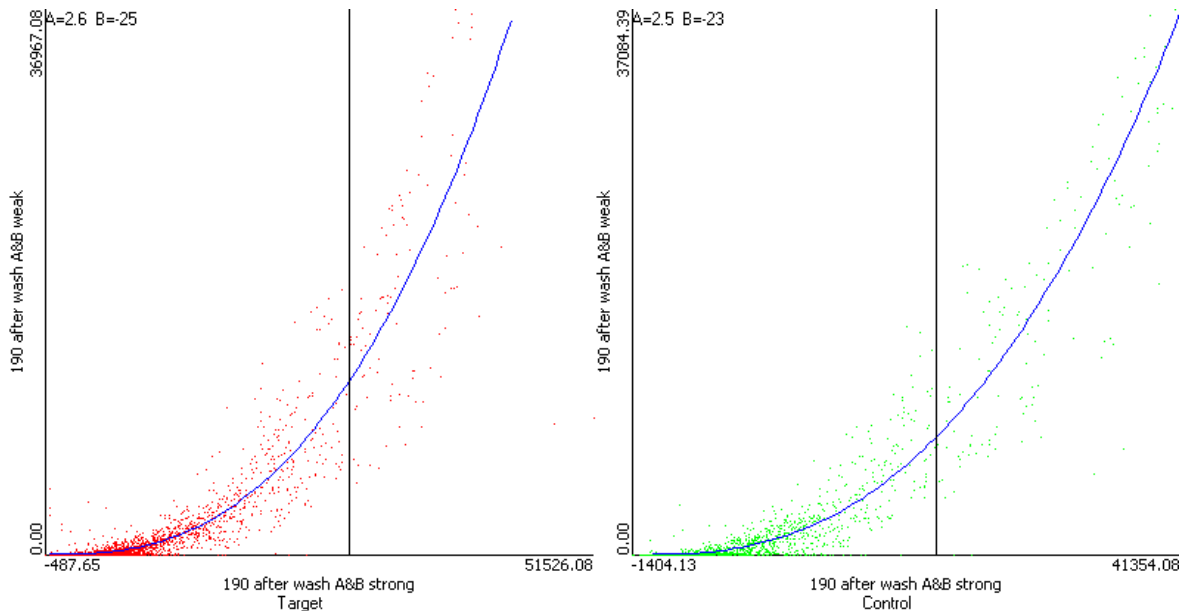
### 6.5.- Double

[This procedure is under patent process, this is a preliminary version in the application]

Fig. 0, Position of the number of replicated points.



Since most *scanners* have the ability to fine-tuning a sensibility level, we have think to take profit of this fact to improve the quality of gene-expression data. In simple terms, a double scan produce two collections of data (for each channel), but this data represent different values for the same spot, thus we need a model and a procedure to found the true value for each spot. However, before solving this problem (found an unique collection of data from two collections) what we need is to establish the relationship between them. This is made by choosing this action. The intensity-intensity view shows the data regression between both *slides*. To launch this action it is necessary, previously, group both slides and set a mark to the slide with the low-sensibility.



**Fig. 0, Data regression in the intensity-intensity view**

As can be observed, the regression produce a typical *gamma*-curve. In this graph, the *slide* obtained with low-sensibility is set in the abscissa axe.



## 7.- Steps: Operation over the states

The state-operations are available in the Step-toolbar.



Fig. 0, The Step toolbar.

From left to right:

- **Ratio Adjusting:** Adjust the ratio zero when this value have changed due to non-linearity in the scan.
- **Scale Adjusting:** Adjust the *slides* scale, when these values are changed due to differences in contrast of sensibility in the *scan*.
- **Filtering:** Remove those spots that do not satisfy a given threshold of intensity.
- **Solving the *Dye-swap*:** Mix up two *slides* performed with the dyes interchanged. Solve the systematic errors due to dyes.
- **Solving replications:** Mix up two spots that measure the same DNA molecule in the same *slide*. This correspond to a *intra-slide* replication.
- **Solving the *double-scan*:** Mix up two *slides* performed with different level of sensitivity in the device.

### 7.1.- Ratio Adjusting

The effect of non-linearity and the different efficiency of dyes produce that intensity in the two channel become different. This difference can be estimated and corrected in this sep-.

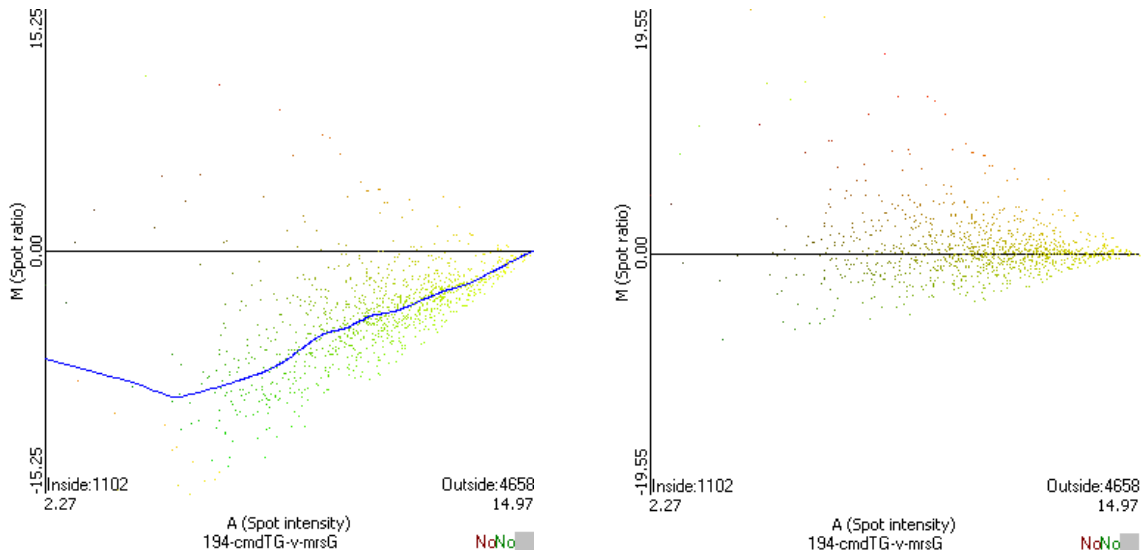


Fig. 0, *Ratio* adjusting. On the left, the input data, and on the right after *thelowess* adjust has been performed.



## 7.2.- Scale Adjust

Different conditions in the *scan* environment, such as temperature and other factor, produce variations in the scan contrast between different *scans*. To correct this irregular contrast, the *ratio* is scaled in the *slides*. The first approach is to use the standard deviation to estimate the contrast, but this method is not robust enough because it take into account the central data as well as the *outliers*. A second approach have been implemented based in the median value.

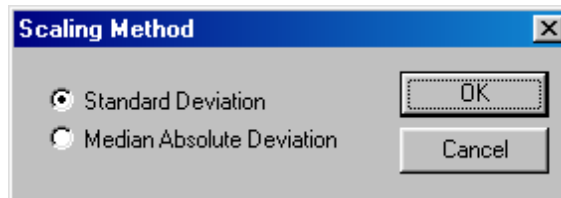


Fig. 0, Dialog box to select the method for scaling.

The boxes diagram allows to visualise the median and the data distribution.

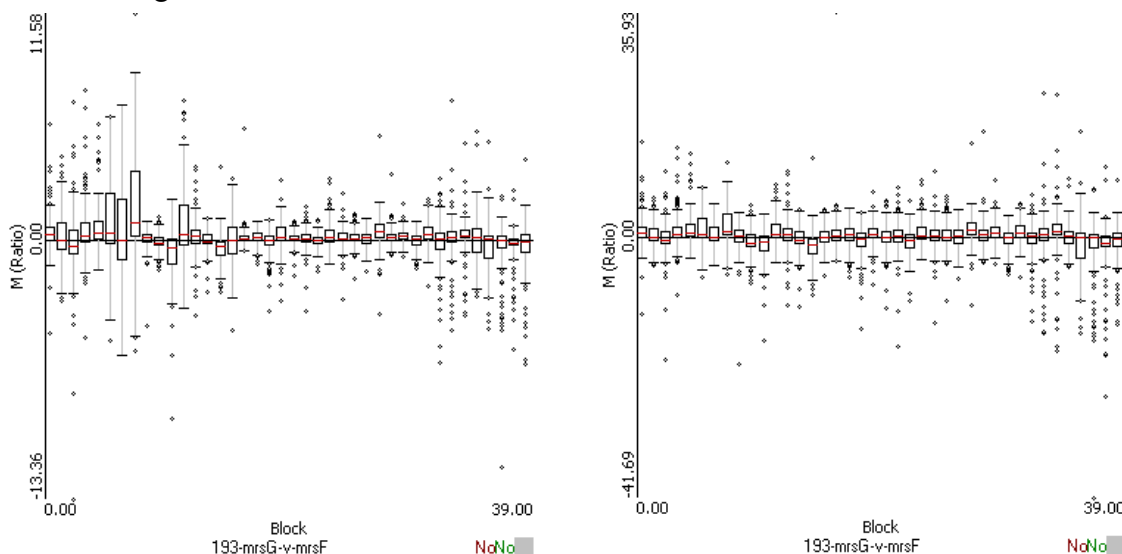


Fig. 0, Scale adjusting. Left, before the adjust, on the right the data distribution after the adjust have been applied (a more uniform distribution of data can be observed).

## 7.3.- Filtering

Due to the quantitation effect, when the intensity in the channels is low, the relative error is high. Moreover, if a given procedure use the error as a quotient, it can produce a completely random result. Thus, it is recommended to remove those points with a very low intensity value in order to avoid this type of error. To proceed to filtering it is necessary to previously establish the thresholds. It is noteworthy to observe that the “units” of the thresholds correspond to the units used in the axis of AM graph.

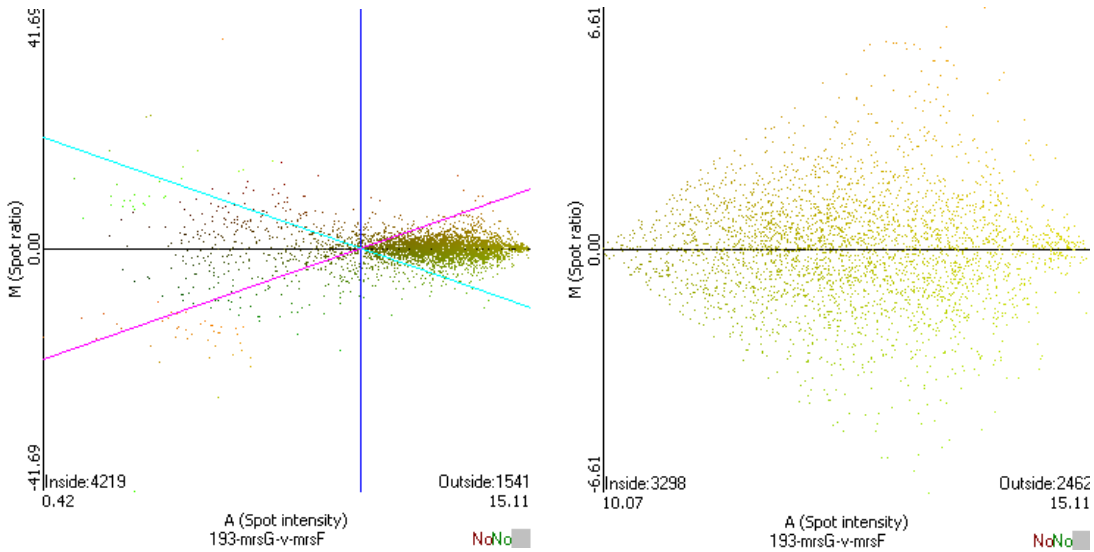


Fig. 0, Filtering spots. On the left, before the filtering—the thresholds are shown. On the right, after the filtering.

### 7.4.- Solving the dye-swap

If the error in the measure is produce for differences in the dye efficiency, it is possible to compensate errors by performing the same experiment with the dyes interchanged. As results we will have two slides that should be mixed to produce only one. To mix up, it is necessary to pair-wise grouping slides that correspond to the same experiment, and set a mark the slide that have the dyes interchanged.

### 7.5.- Solving replications

Some random errors can not be solved, but some methods can reduce their effect. These methods use a replication of the measures, and proceed to estimate the real value. Two different metjods have been implemented, and the dialog box is used to choose one of them..

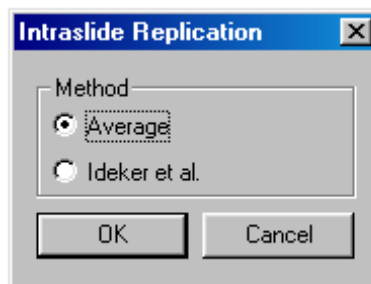


Fig. 0, Dialog box to choose the method to sdve replicated experiments (intra-slide).



- **Average:** Based on the average of spots. This method perform well when the number of replications is low and there is not much a priori information.
- **Ideker et al.:** Performs an estimation of the noise. But much more information is needed: first, a high number of replicated spots and, second, it assume the distributions are normal.

## 7.6.- Solving the double-*scan*

Scanning the same chip with different saturation level (sensitivity level) is equivalent to have a *inter-slide* replication. However, this replication needs an evaluation of the real value of the intensities. This estimation is obtained by regression. In this way the final value for each spot correspond to the closest value to the regression curve (this is a first approach, we are currently working in a more elaborated procedure). To perform this task, *slides* must be grouped by pairs, and the slide obtained with low sensitivity needs to be marked.

*[This procedure is under patent process, thus it is not available in this application]*



## Annexes

### Annexe A, File format for slide data

The input slide data files have a simple tabular format (i.e. Excel can be used to produce a tab file). Columns are the column separator, and new-line the row separator. The first line contains the column labels, and the rest of lines, the values. Each line correspond to one point in the *slide*. Some DNA-array software produce a final statistical analysis of data, that *must* be deleted to allow **PreP** to work properly.

	A	B	C	D	E	F	G	H	I
1	#ORF	x	y	Control-sig	Control-bkg	Target-sig	Target-bkg		
2	BG10065	1	1	382	328	250	222	dnaA, dnaH, dnaJ, dnaK	
3	BG10066	2	1	377	328	257	222	dnaN, dnaG, dnaK	
4	BG10067	3	1	2331	328	2640	222	yaaA	
5	BG10068	4	1	1893	328	1607	222	recF	
6	BG10069	5	1	1213	328	1048	222	yaaB	
7	BG10070	6	1	1914	328	1821	222	gyrB, novA	
8	BG10072	7	1	917	328	621	222	yaaC	
9	BG10073	8	1	2123	328	2268	222	guaB, guaA, gnaB	
10	BG10074	9	1	1785	328	869	222	dacA	
11	BG10075	10	1	2714	328	1408	222	yaaD	
12	BG10076	11	1	7910	328	8427	222	yaaE	
13	BG10077	12	1	383	328	262	222	serS	
14	BG10078	13	1	390	328	274	222	dck, yaaF	
15	BG10079	14	1	486	328	341	222	dgk, yaaG	
16	BG10080	15	1	568	328	388	222	yaaH	
17	BG10081	16	1	475	328	368	222	yaaI	
18	BG10082	17	1	669	328	469	222	yaaJ	
19	BG10083	18	1	2164	328	1812	222	dnaX, dnaH, dna-8132	
20	BG10084	19	1	8010	328	5999	222	yaaK	
21	BG10085	20	1	2843	328	2238	222	recR, recM	

**Fig. 0, Excel visualization of a slide file.**



## Annexe B, Output File Format (*engine* format).

This is also a tabulated format and Excel can also be used for visualization and edition. A “tab” is used between columns and a new-line between rows. The simplest *engine* format does not need labels, thus, only the data are necessary:

	A	B	C
1	1	16	72
2	123	15	1
3	151	15	23
4	32	516	53

**Fig. 0, The simplest *engine* format is composed only by data.**

Missing values are filled with a special value (not numeric) called NaN (Not a Number).

	A	B	C
1	NAN	16	72
2	123	15	UNK
3	151		23
4	32	516	53

**Fig. 0, Missing values in red.**

Labels help to identify the meaning of each data. This type of information is named “*metadata*”(data about data). It is possible to insert labels for rows and columns, and each label have a name. Next picture shows the label position:

	A	B	C	D	E
1		Ctag1Name	Ctag1Val1	Ctag1Val2	Ctag1Val3
2		Ctag2Name	Ctag2Val1	Ctag2Val2	Ctag2Val3
3	Rtag1Name				
4	Rtag1Val1		1	16	72
5	Rtag1Val2		123	15	1
6	Rtag1Val3		151	15	23
7	Rtag1Val4		32	516	53

**Fig. 0, Label distribution in the data file.**

- **Red:** Column label names. This label name links all the label-values on the right (in yellow).
- **Yellow:** The column label-values. Each label-value is linked to (describe) the column values below it.



- **Green:** Row label-names. This label name links all the label-values below it (blue).
- **Blue:** Row label-values. Each label-value is linked to (describe) the row values on the right it.
- 

Finally the table structure have the following format:

	A	B	C	D	E
1		Ctag1Name	Ctag1Val1	Ctag1Val2	Ctag1Val3
2		Ctag2Name	Ctag2Val1	Ctag2Val2	Ctag2Val3
3	RTag1Name				
4	RTag1Val1		1	16	72
5	RTag1Val2		123	15	1
6	RTag1Val3		151	15	23
7	RTag1Val4		32	516	53

**Fig. 0, Data Table structure.**

»S

These cells separate the data and the labels.

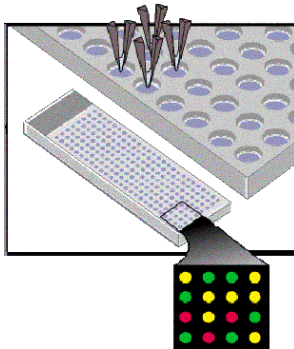
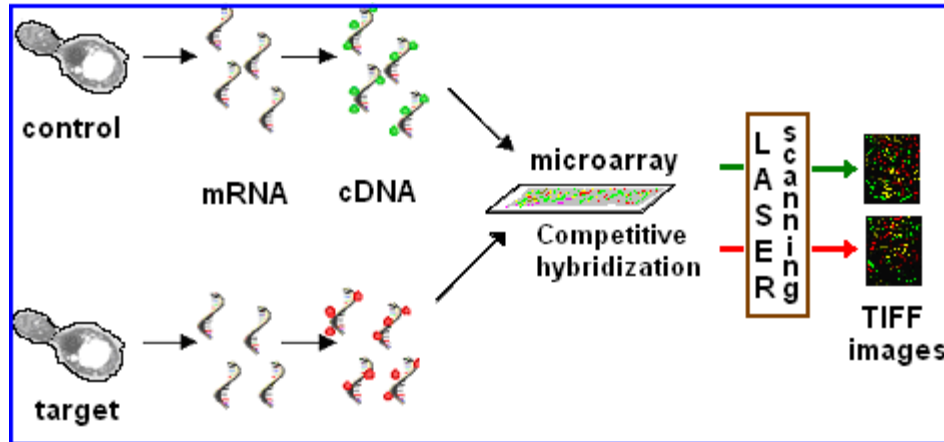
- **Cyan:** Empty cells are necessary over and left the labels. These cells separate the labels between them.

Note: when using Excel it is necessary to sep in mind the “local” numeric representation (*engine* format does not use the comma as thousands separator, and decimals are separated by a dot).



## Annex C: The initial source of data

The initial source of raw-data in the EF-project will be formed by cDNA array measurements on a large collection of *L.Lactis* strains under different experimental conditions. For each experiment, competitive hybridization with sample cDNA (labelled with fluorescent tags) will be recorded on pairs of image files (16-bit TIFFs), each image corresponding to one of the dyes in the particular experiment (for details see Annex 2).



Cells are grown in the control and experimental mediums (bacteria, such as *L.Lactis* are single-celled organisms). To ensure reproducibility across experiments, the cells are grown to a fixed *optical density* (density of cells in the culture) in the exponential phase of growth, at which point they are harvested. Control and target mRNAs are reverse transcribed into cDNA and differentially labelled with the dyes Cy3 (Green) and Cy5 (Red). The two cDNA samples are competitively hybridized to a DNA array. Subsequently, the DNA array is scanned using two different channels (scanning the Green and Red fluorescent signals), yielding two images..

Image analysis is applied (using the spot-signal quantification software ArrayPro) to extract measures of the red and green fluorescence intensities for each spot on the array. The image processing procedure should produce the following (minimal) data : (a) gene-position in the image (x,y); (b) estimate location of spot centers (Xcoord, Ycoord); (c) pixel classification (signal and background) and (d) for each spot on the array and each dye: (signal and background intensities and quality measures (absent, marginal or valid)

As results of this step, the data will look like as shown in Table 1 (data corresponds to Kobayashi experiments on *B.subtilis*)



#ORF	x	y	Xcoord	Ycoord	Control-sig	Control-bkg	Target-sig	Target-bkg	geneName
BG10065	1	1	50	52	938	189	725	249	dnaA, dnaH, dnaJ, dnaK
BG10066	2	1	102	51	2692	189	2253	249	dnaN, dnaG, dnaK
BG10067	3	1	153	49	958	189	444	249	yaaA
BG10068	4	1	201	50	6703	189	2533	249	recF
BG10069	5	1	255	51	496	189	327	249	yaaB
BG10070	6	1	302	50	4337	189	2017	249	gyrB, novA
BG10119	60	1			6385	189	4128	249	spoVT, yabL
BG10120	61	1			967	189	823	249	yabM
BG10121	62	1			8740	189	6273	249	yabN
BG10122	63	1			1199	189	1364	249	yabO
BG10123	64	1			960	189	849	249	yabP
BG10124	1	2			436	189	796	249	yabQ
BG10125	2	2			14823	189	12676	249	divC, divA
BG10126	3	2			722	189	524	249	yabR
BG10127	4	2			323	189	388	249	spollE, spollH, spollK
BG11465	62	63			415	189	389	249	phrG, yycL
BG10932	63	63			2148	189	11056	249	rocF
BG10933	64	63			1062	189	6838	249	rocE
BG10722	1	64			1106	189	7821	249	rocD
BG10723	2	64			437	189	636	249	rocR
BG10051	59	64			467	189	358	249	yyaE
BG10052	60	64			285	189	372	249	yyaD
BG10053	61	64			204	189	296	249	yyaC
BG10054	62	64			962	189	909	249	spoDJ
BG10055	63	64			1710	189	1424	249	soj
BG10056	64	64			351	189	390	249	yyaB

Table 1. Data from image analysis procedures



## Annexe D: Image Analysis issues

This annexe is devoted to image processing general issues applied to gene expression data. This is aimed only as general information about the way in which data are obtained. Extensive literature is available for details.

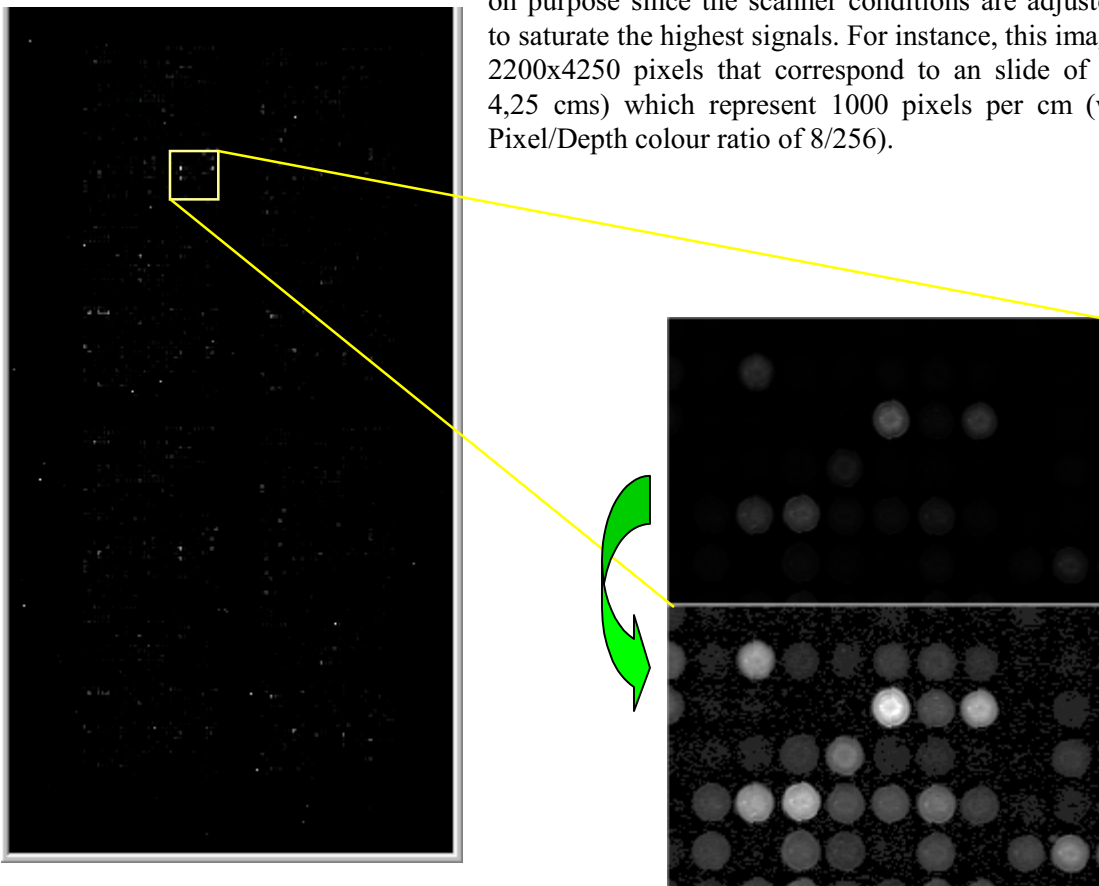
### Experiment Procedure

Let suppose being working on the next experiment:

- Two kinds of RNA samples has been prepared from B.Subtilis cells grown on different conditions to be compared
- The same amounts of the two RNA samples were subject to reversed transcription separately to obtain respective cDNAs, and labelled with Cy3 and Cy5 respectively.
- The two above differentially labelled cDNAs were mixed, and hybridised onto a slide glass carrying microarray of spots of the PCR products specific to respective B.subtilis ORFs.
- After completion of the hybridisation processes the slide glass was scanned through two channels for Cy3 and Cy5 to obtain the fluorescence signals, thus producing two image files from one slide glass.

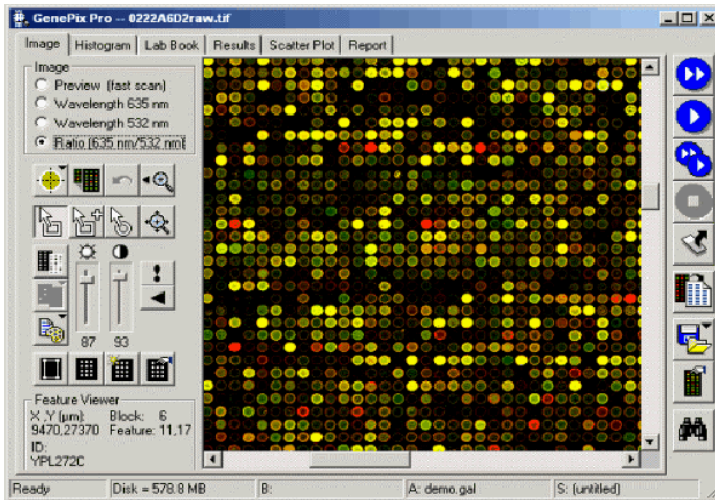
As described above, one experiment provides us with twin images files; one is through the Cy3 channel and the second from the Cy5 channel. Each of the images are produced by scanning a whole one slide glass

In general the images looks monochrome and can be rather dark as we ll (see figure). This is made on purpose since the scanner conditions are adjusted not to saturate the highest signals. For instance, this image is a 2200x4250 pixels that correspond to an slide of (2.2 x 4,25 cms) which represent 1000 pixels per cm (with a Pixel/Depth colour ratio of 8/256).





For visual inspection needs, the brightness/contrast can be changed to make the spots more obvious (avoid saving the new images, otherwise the final results shall be altered). Third party software is used to overlay the twin images precisely and to obtain the quantitative data of the signal intensities found on exactly the same area spot on the two overlaid images. So far, it is experienced in most cases that it is possible to normalize the obtained data, in good agreement with expected outcomes, on the assumption that the sum total of the signal intensities of all spots on each of the twin images (representing the overall transcriptome) could be constant.



In this way, the initial raw data from a cDNA microarray experiment is formed by pairs of image-files, in general 16-bit TIFFs, one for each of the dyes. Image analysis is required to extract measures of the red and green fluorescence intensities for each spot in the array. These measurements are obtained using proprietary software, as in the image above in which the GenePix Pro analyser is used.

The following steps are the most frequently used:

- 1.- Estimation of the spot centre locations (left figure)
- 2.- Segmentation, aimed to classify pixels as foreground (signal) or background (noise)
- 3.- Information extraction for each spot on the array and for each dye: a) signal intensities (b) background intensities and (c) quality measures

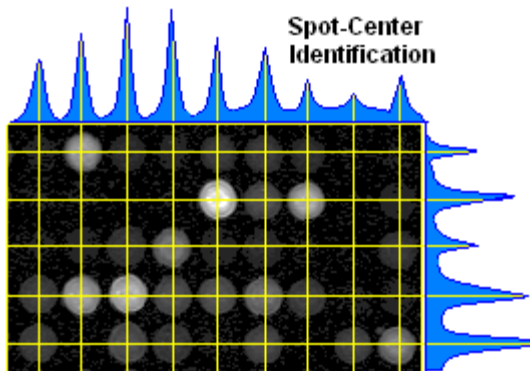
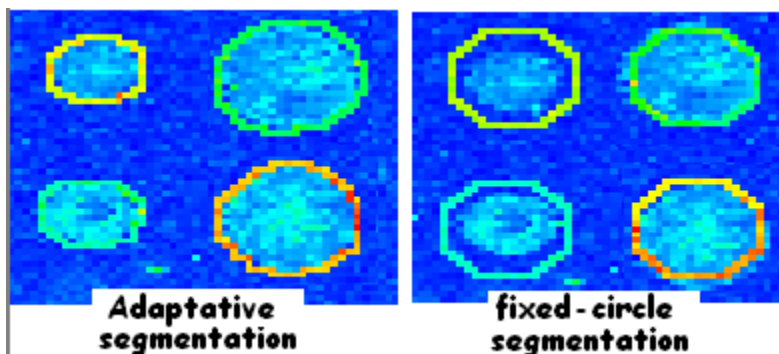


Image segmentation can be addressed in a multitude of different ways. The most frequent are: adaptive and fixed circle segmentation, since spots usually vary in shape and size



Adaptive segmentation methods often make use of seeded region growing algorithms that extract connected component around a seed region. Thus given a starting pixel (the seed) with

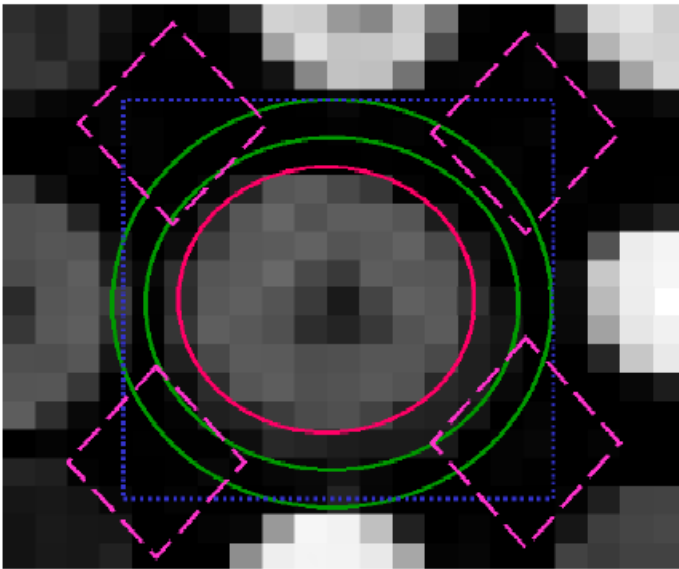
black value (in general a grey level with a value up to a given threshold). The region growing algorithm finds all pixels that are black and connected to it.

Colloquially, the region growing algorithm can be described as follows: each pixel in an image has 8 neighbours (for example, the pixel (4,6) has the pixels at locations (4,5), (4,7), (5,5), (5,6), (5,7),



(3,5),(3,6), (3,7) as neighbours. Not all the neighbours are black, thus the algorithm only consider the neighbouring pixels that are black. There are two ways to proceed: depth first or breadth first. In depth first search, you recursively follow one neighbour of a pixel. In breadth first search you process all the neighbours of one pixel before proceeding to the next one.

As can be see, this procedure requires the input of *seeds*, either individual pixels or groups of pixels which control the formation of the regions into which the image will be segmented. In automatic segmentation the seeds can be based on fitted foreground and background grids, and the decision to add a pixel to a region is based on the absolute grey -level difference of that pixels intensity and the average of the pixel value in the neighbouring region. (done on combined red and green images)



---- GenePix    ---- QuantArray    ---- ScanAnalyzer

foreground/background ratio

- Uniformity: variation in pixels intensities and ratios of intensities
- Morphology: area, perimeter, circularity

Slide quality

- Percentage of spots with no signal
- Range of intensities
- Distribution of spot signal area, etc.

To estimate the local background intensities, the image is tested with a structuring element (i.e. a square with side length about twice the spot to spot distance, as shown on the left). Over this area opening morphological filters are applied (erosion followed by dilation). The eroded (or the dilated) value at a pixel  $x$  is the minimum (maximum) value of the image in the window defined by the structuring element when its origin is at  $x$ . This procedure is done separately for the red and green images, and produce an image of the estimated background for the entire slide.

Quality measures are referred to the following issues:

- Brightness:

The most important question posed in this aspects is “how to use quality measures in subsequent analysis?”.

*throughout data normalization procedures, aimed to identify and remove sources of systematic variation in the measured fluorescence intensities, other than differential expression*