

PreP: Gene-Expression Data Pre-Processing Tool

Jorge García de la Nava¹, Sacha van Hijum² and Oswaldo Trelles^{1,*}

¹Computer Architecture Department, Universidad de Málaga, 29080, Málaga, Spain

²University of Groningen, Molecular Genetics, The Netherlands.

GUIDED TOUR

This manual is a work in progress
by Jorge García de la Nava
gdl@ac.uma.es

Biological side is cover by
Sacha van Hijum
S.A.F.T.van.Hijum@biol.rug.nl

Project Director
Dr. Oswaldo Trelles
ots@ac.uma.es

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

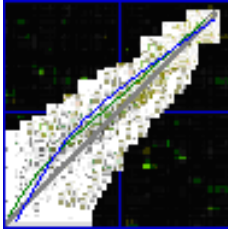
Bioinformatics and Genomics Laboratory, University of Málaga
<http://chirimoyo.ac.uma.es/bitlab/index.html>

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



PreP: Gene-Expression Data Pre-Processing Tool

Jorge García de la Nava¹, Sacha van Hijum² and Oswaldo Trelles^{1,*}

¹Computer Architecture Department, Universidad de Málaga, 29080, Málaga, Spain

²University of Groningen, Molecular Genetics, The Netherlands.

GUIDED TOUR

This document contains a Guided Tour through the **PreP** platform and it was created for training purposes with respect to the system options and analysis possibilities. It is not intended for training about the biological interpretation of the results.

Original data used in this demo correspond to Kobayashi et al. (2001), “Comprehensive DNA microarray analysis of Bacillus subtilis two-component regulatory systems”. J Bacteriol. 183(24): 7365-70” and it is available in the web at:

http://www.genome.ad.jp/dbget-bin/get_htext?Exp_DB+-n+B

They can be obtained in **PreP** format directly from the home page at:

<http://www.engene.cnb.uam.es/downloads/kobayashi.dat>

Contents

- 1.- Guide Scheme 3
 - 1.1.- Steps..... 3
 - 1.2.- Data Set..... 3
- 2.- Load Step 3
- 3.- Adjust and Ratio correction 7
 - 3.1.- Block Selection 7
 - 3.2.- Adjust..... 8
 - 3.3.- Apply the Correction 9
- 4.- Scaling 10
- 5.- Filtering..... 12
 - 5.1.-Threshold selection 12
 - 5.2.- Filtering..... 13
- 6.- Closing words 14

Note: this file contains a complete demo. However, it will be modified and extended as new algorithms will be available, and as new experience with the system will be reported. Please, submit any recommendation or suggestions to our webmaster from the **PreP**-home page, or

directly to gdl@ac.uma.es



PreP: Guided Tour

1.- Guide Scheme

1.1.- Steps

PreP is oriented to apply different algorithms on the same data set. This guided tour will shown each of these procedures, in the following order:

- a) Load Step: is the first and the only mandatory step. In this step the structure of data is established.
- b) Adjust and ratio correction
- c) Data scaling
- d) Thresholding and filtering
- e) *dye-swap* replicates
- f) *Intraslide* replication
- g) Save results.

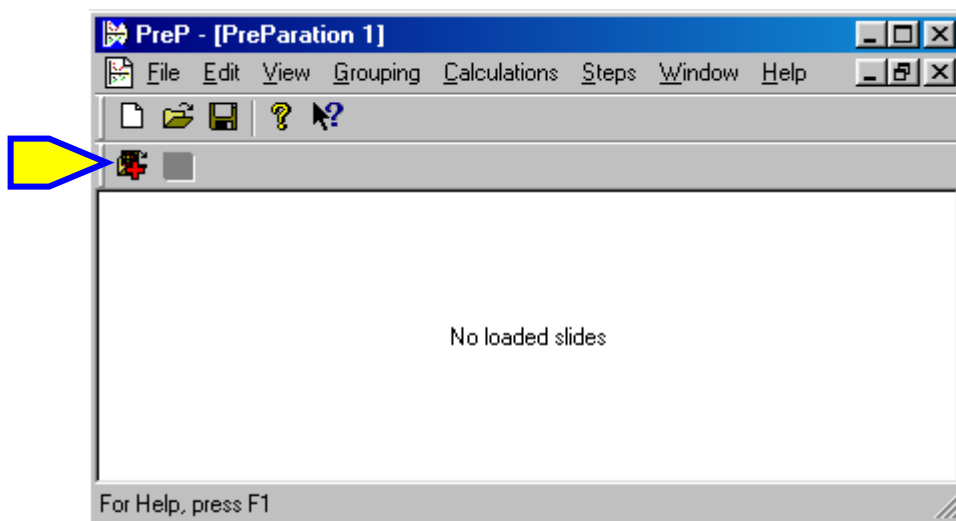
1.2.- Data Set

Test data sets are available in our web site (<http://chirimoyo.ac.uma.es/bitlab>), they correspond to DNA array experiments with bacillus subtilis genes, also available at the KEGG Expression database (http://www.genome.ad.jp/dbget-bin/get_htext?Exp_DB+n+B). PreP recognise the “.slide” file extension. Thus, we recommend to rename the files if necessary.

2.- Load Step

2.1.- Browsing the file

From the initial screen, the “load step icon” can be used to launch the load procedure. A new slide will be added to the empty document. A “file box dialog” is used to allow browse and specify the file. Try with “ex0000260.slide”.



Information about the loaded file is shown in the window.

The screenshot shows the 'PreP - [PreParation 1]' window with a menu bar (File, Edit, View, Grouping, Calculations, Steps, Window, Help) and a toolbar. Below the toolbar is a table with the following data:

Spots	4005
Name	ex0000260
#ORF	-
x	-
y	-
Control-sig	-
Control-bkg	-
Target-sig	-
Target-bkg	-

At the bottom of the window, it says 'For Help, press F1'.

First two rows correspond to the number of spots and the file name. Following rows correspond to the “row labels”. Each column represent a different file or slide (in this case only the ex0000260 has been loaded). Observe that the file name have a “red” colour, meaning that no functionality has been established yet.

To visualize the *slide* data, click the left button over the slide filename.

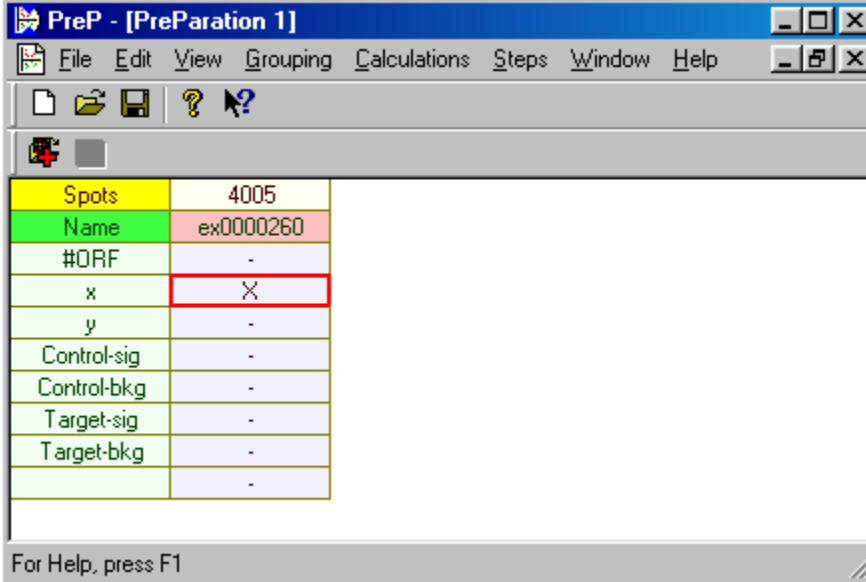
The screenshot shows the 'PreP - [PreParation 1]' window with a menu bar and toolbar. Below the toolbar is a table with the following data:

#ORF	x	y	Control-sig	Control-bkg	Target-sig	Target-bkg	
BG10065	1	1	305	209	169	73	dnaA, dnaH, d...
BG10066	2	1	262	209	88	73	dnaN, dnaG, d...
BG10067	3	1	1093	209	1544	73	yaaA
BG10068	4	1	1813	209	1979	73	recF
BG10069	5	1	855	209	849	73	yaaB
BG10070	6	1	1306	209	1466	73	gyrB, novA
BG10072	7	1	417	209	219	73	yaaC
BG10073	8	1	807	209	475	73	guaB, guaA, g...
BG10074	9	1	511	209	1256	73	dacA
BG10075	10	1	2492	209	3937	73	yaaD
BG10076	11	1	5278	209	9858	73	yaaE
BG10077	12	1	294	209	159	73	serS
BG10078	13	1	287	209	152	73	dck, yaaF
BG10079	14	1	411	209	464	73	dgk, yaaG
BG10080	15	1	384	209	183	73	yaaH

At the bottom of the window, it says 'For Help, press F1'.

To go back to the previous view, click the mouse over the window. Additional functions are available with the right mouse button (over the *slide* filename)

2.2.- Assigning functionality to labels

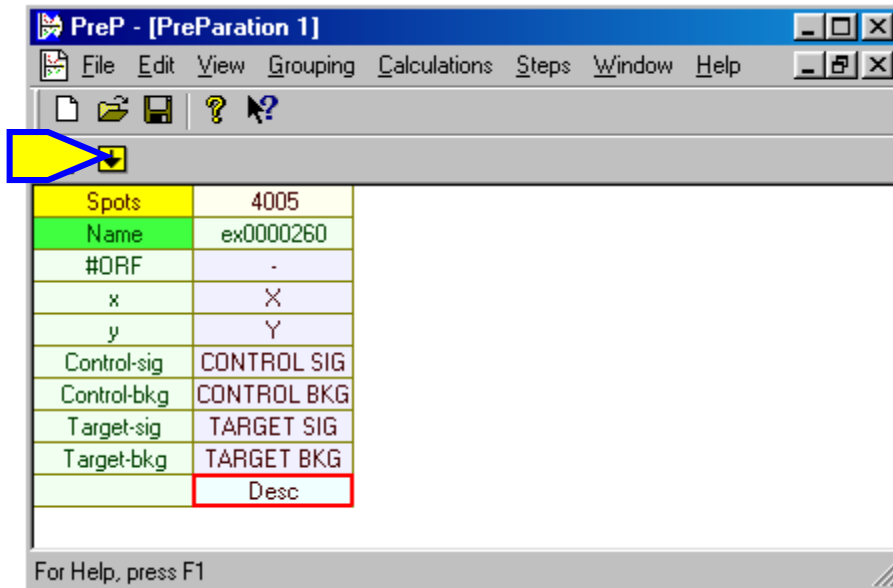


Using the cursor keys move the red-box to the “x” row and press the “X” key to assign the function “coordinate X” to the label “x”.

In the same way, the functionalities shown in the table can be assigned:

Label	Function	Hot Key
Y	Coordinate Y	Y
Control-sig	Control signal	O
Control-bkg	Background in control signal	N
Target-sig	Target signal	A
Target-bkg	Background in target signal	R
<empty>	Description	<Enter>

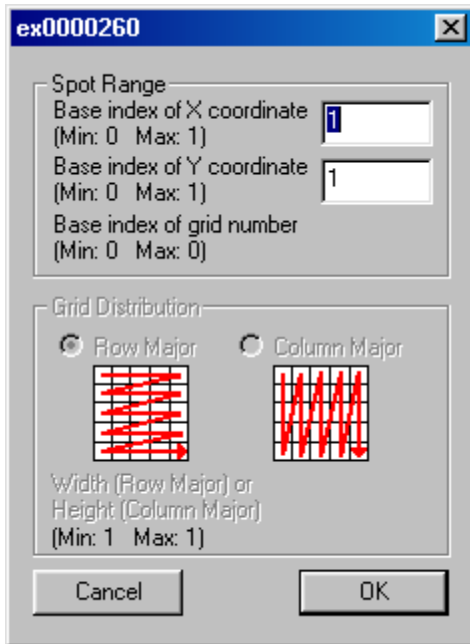
The label for the last row is empty and correspond to gene-description. We should introduce a name to better identify it. Introduce “Desc”:



Now, the slide-name is green, meaning that functionalities has been assigned and PreP is ready for the next step. Push the “next” icon to end the load step. PreP will analyse the file.

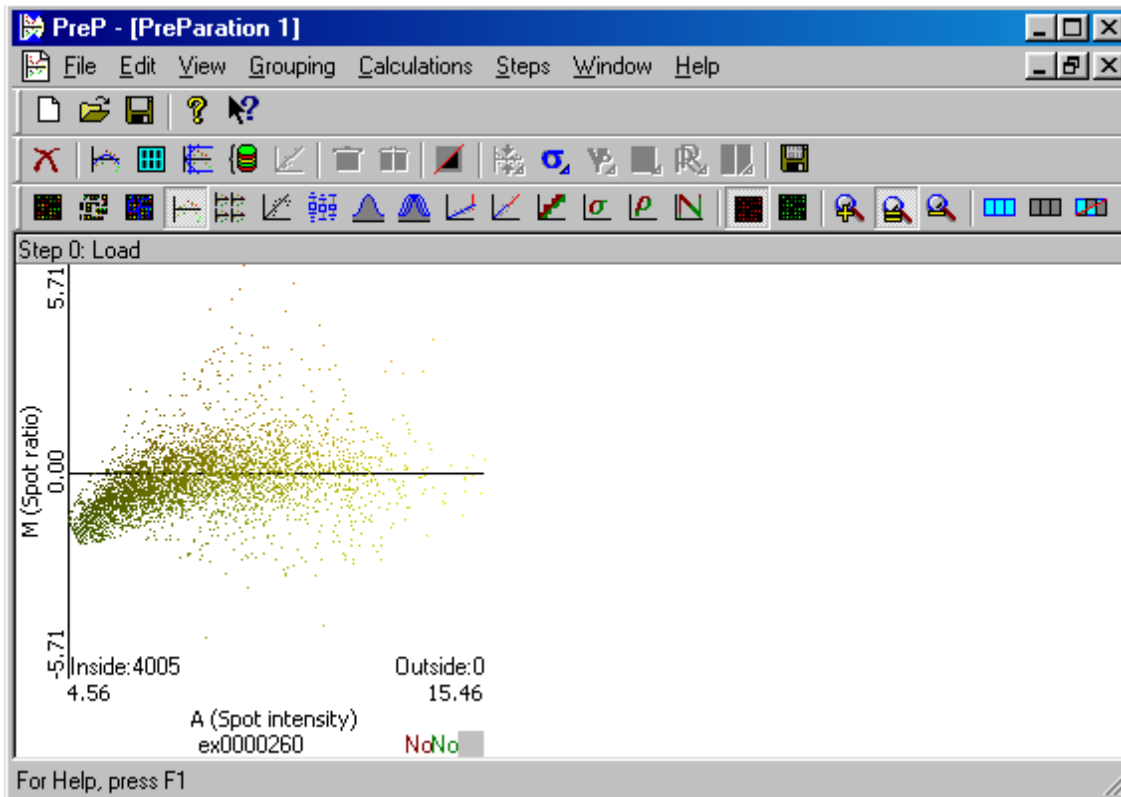


2.3.- Slide structure



PreP is not able to know in advance the *slide* file organization. Next dialog box is designed to complete this information. In our case the default values are enough.

The load step has been completed, thus, PreP compute the AM graph and displays it.



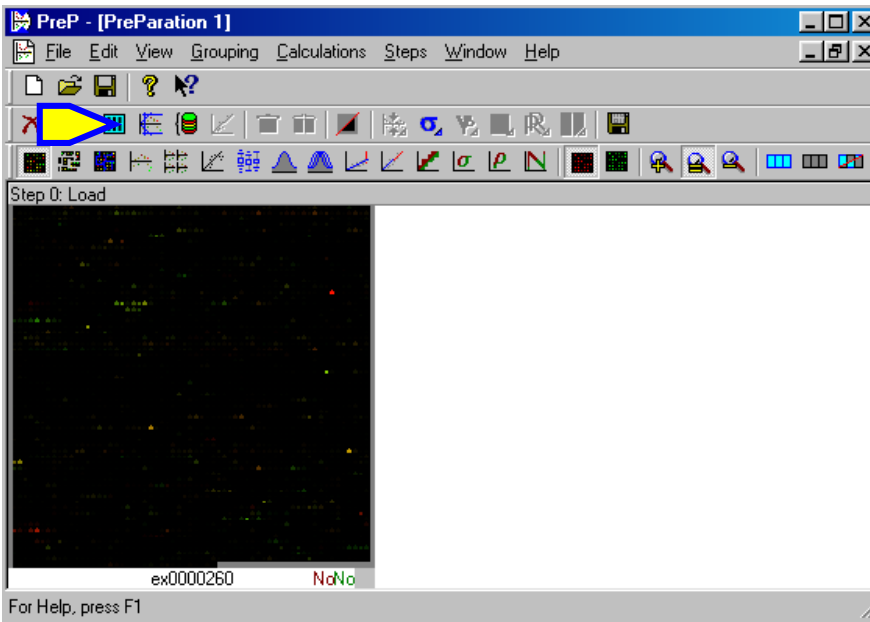


3.- Adjust and Ratio correction

Learning from the “ex0000260” *slide* (see KEGG documentation) we will observe that under the specific experimental conditions, the DegU regulon will be over-expressed. This means that most of the other genes will maintain the same expression level, and only those genes related to DegU will be differentially expressed. Most genes should reflect a ratio zero and should be over the abscissas axe, but this is not the case. However, measurement errors are present and we should try to remove it by ratio compensation.

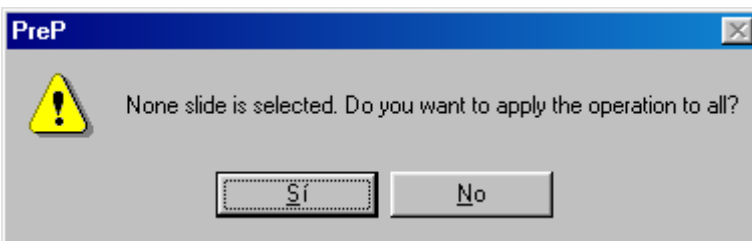
3.1.- Block Selection

Because errors could be affected by the spatial disposition of the spots, we are going to make the adjust by blocks. First we proceed by switching the visualization mode. Click on the icon .

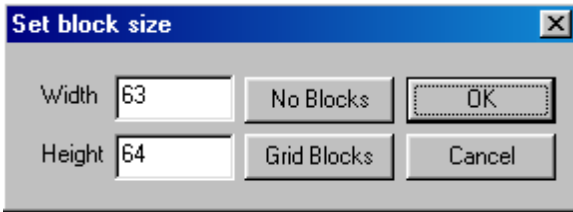


The visualization mode change to:

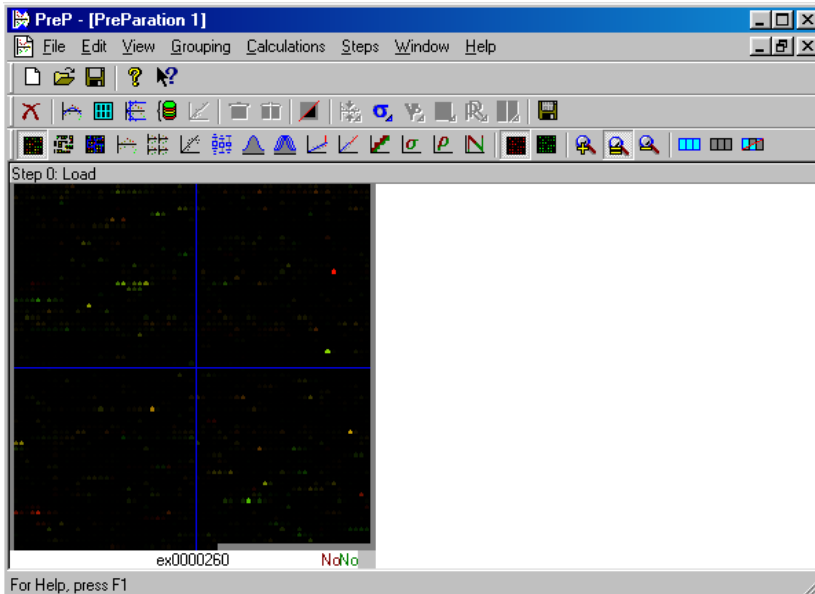
Click on the marked icon to start the “split the slide in blocks” operation.



Apply the operation to all the *slides*.



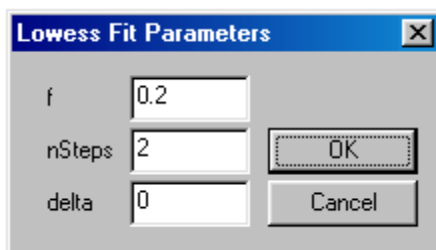
A block size is requested. The (block size) default value correspond to all the slide , or to the grid size if this functionality was specified during the load step (in this example we did not use this option).



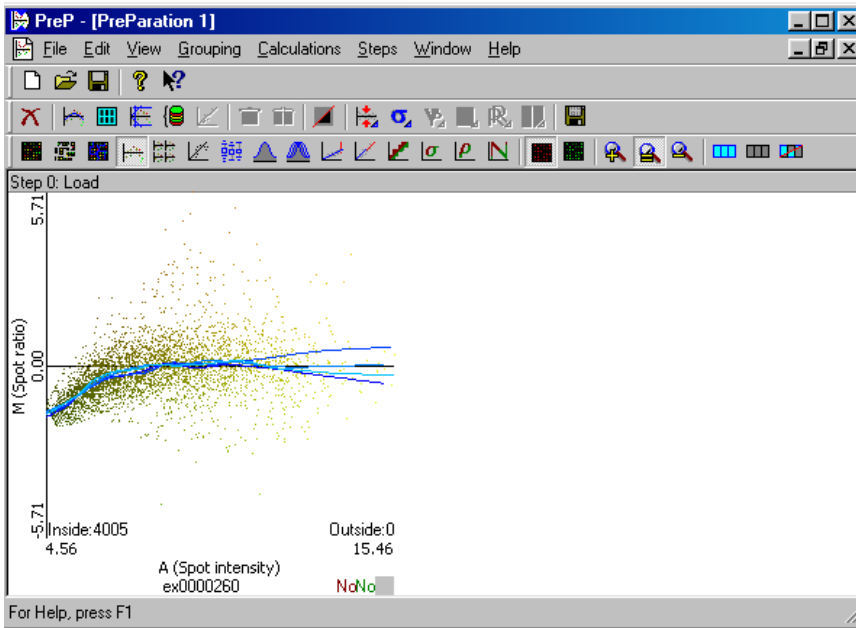
Lets divide the *slide* in fourth blocks, thus we use a 32x32 size (replacing the 63x64 default value). The view will show the borders of each block.

3.2.- Adjust

Now we apply the adjust procedure over each block. The adjust procedure will estimate the error or measurement variations of the rations. Click on the “Adjust icon”:



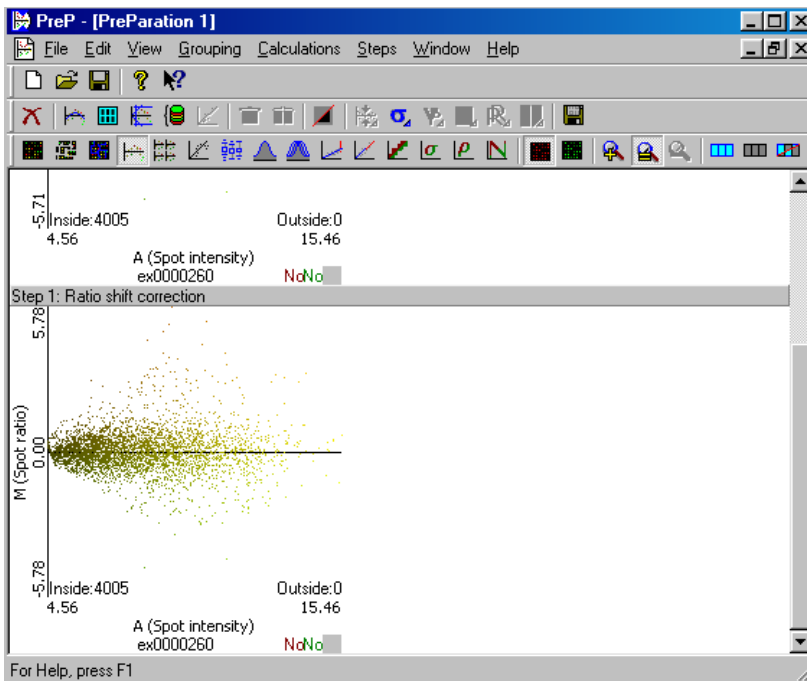
A dialog box will request the adjusting parameters. The procedure implement the “lowess” algorithm. Accept the default values:



Using the AM graph will shown us the adjust curves for each block.:

3.3.- Apply the Correction

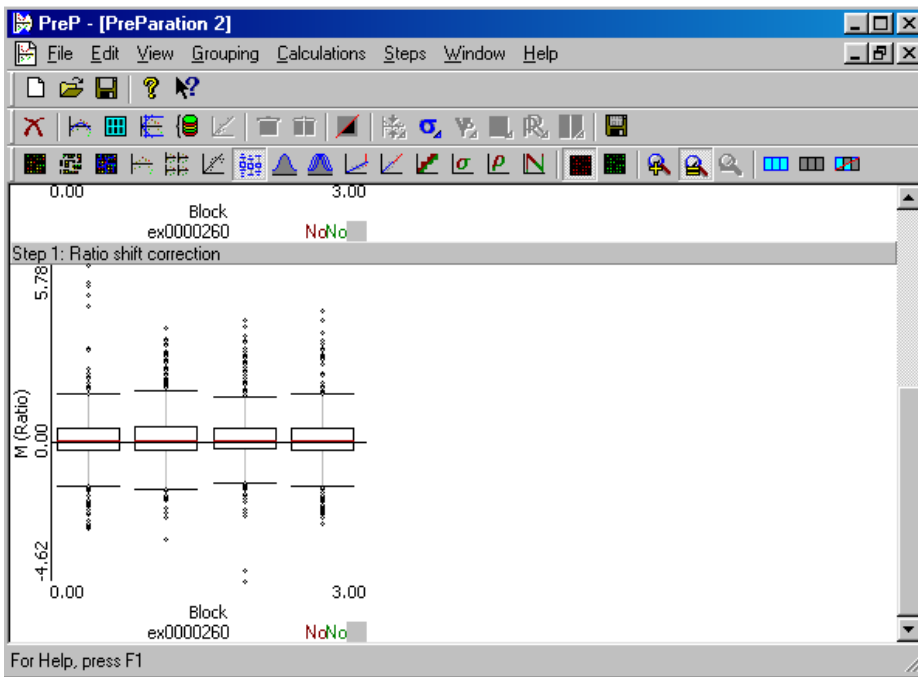
Once obtained the adjust curves we are ready to apply the correction procedure. Click on the icon.



A new state has been produced as result of the correction procedure. Compare this data distribution against the original to observe the effect of the lowess procedure..

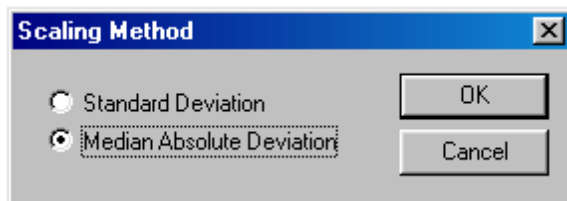
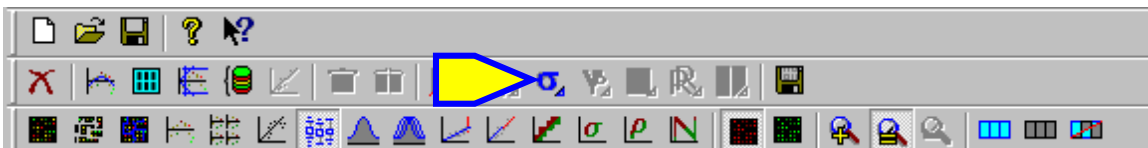
4.- Scaling

Another typical source of error are the contrast variations (non-linearities and different conditions when measuring). Scaling is the technique used to remove differences produced by contrast (in the same slide or between different slides). To visualize the range of data variation use the “boxes graph”, by clicking the corresponding icon.



The view has the following aspect:

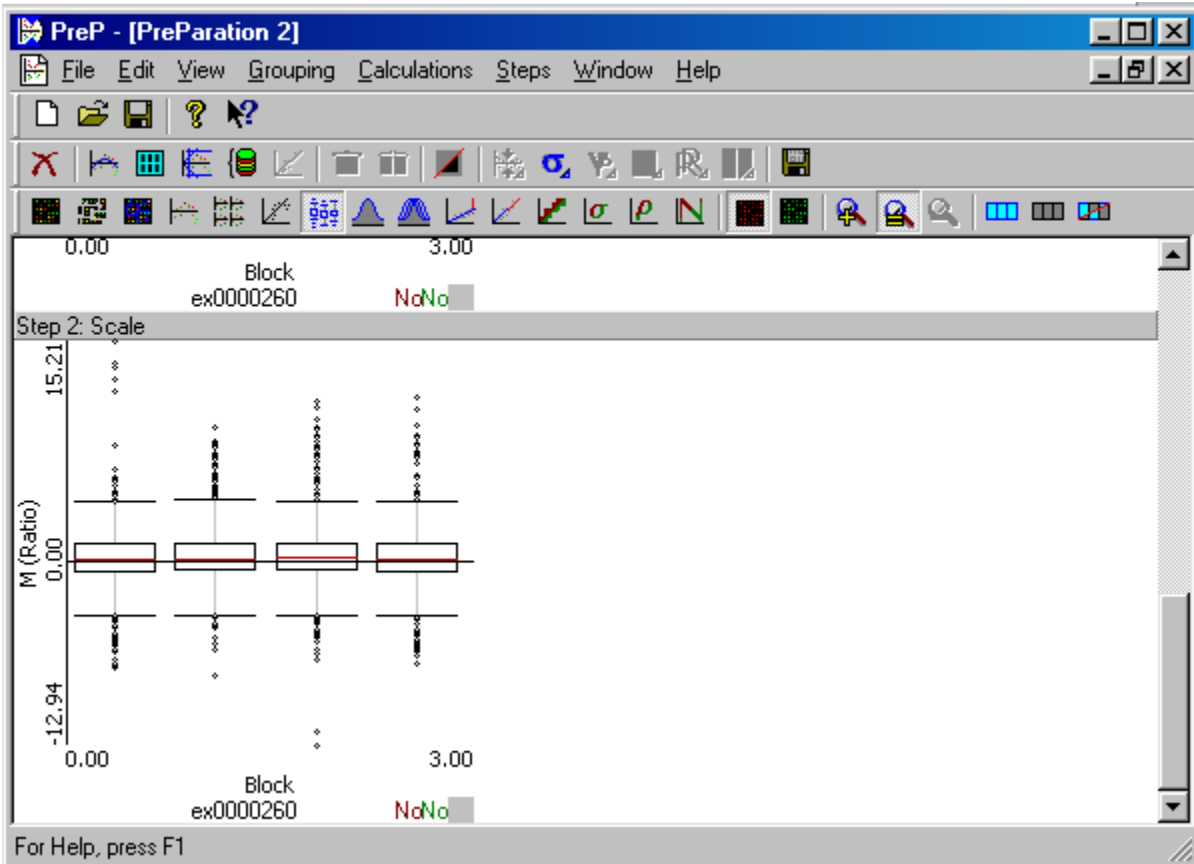
Now, choose the scaling option to make more homogeneous the range of data in the different blocks. Click the icon.



Two different methods are available. Lets use the “median” as the more robust one.



A new state is produced as result of the scaling procedure. When comparing with the previous state a more regular data distribution can be observed.





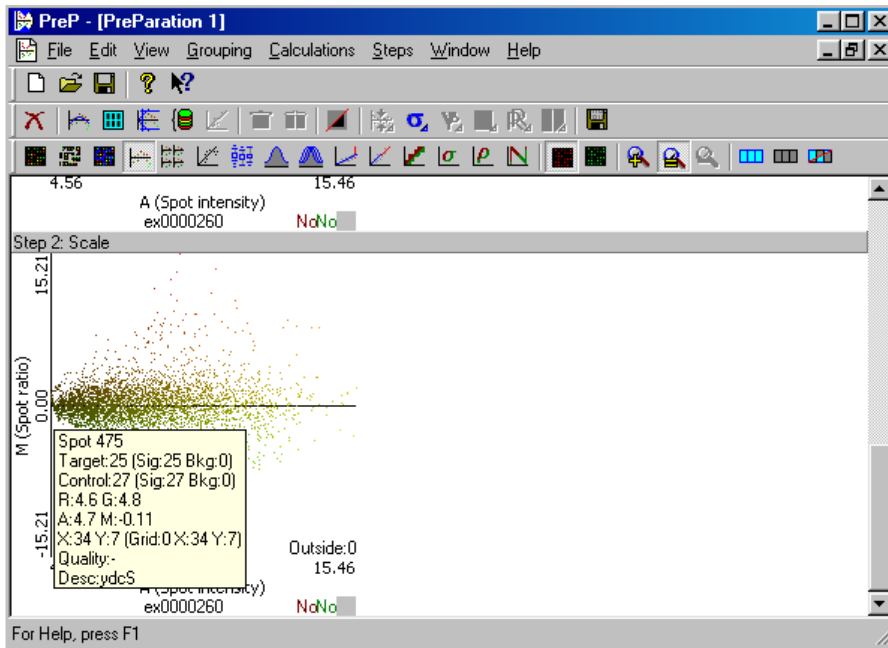
5- Filtering

Technological issues in the scan produce a greater relative error in the low intensities range. Moreover, computing the *ratio*, amplify and propagate this error. The solution is remove the low intensity values by applying a filtering procedure.

5.1.-Threshold selection

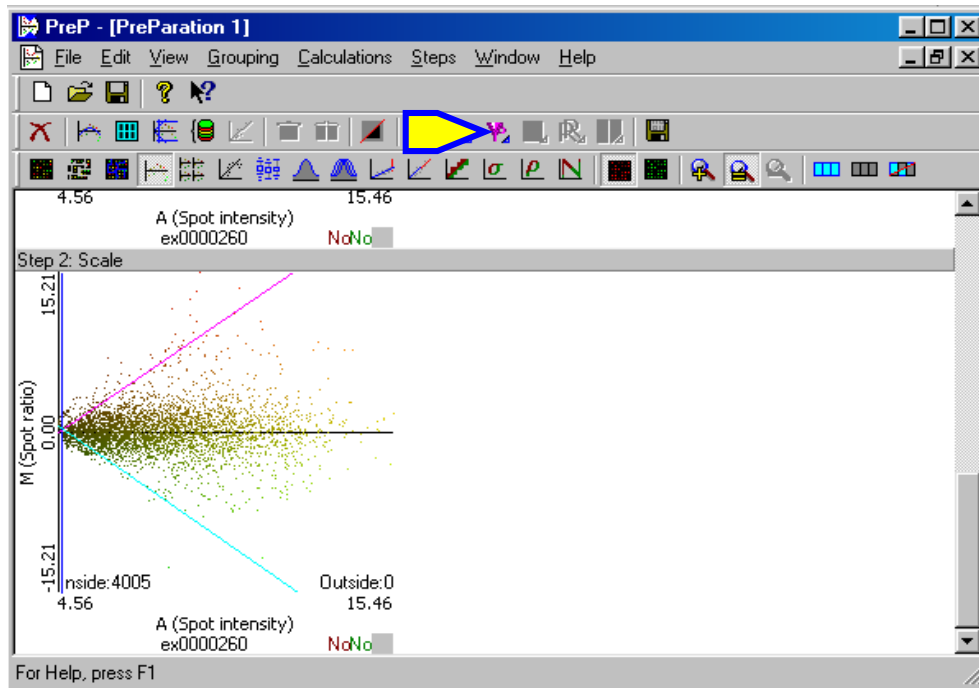
One difficult task is choosing an adequate threshold. Using the mouse over the points we wish to remove we will obtain detailed information about this point. To filter the low intensity points we

should concentrate on the points on the left. In particular, we are interested on the intensities values on both channels and the total intensity. This values correspond to the R, G and A labels (values 4.6, 4.8 and 4.7.)



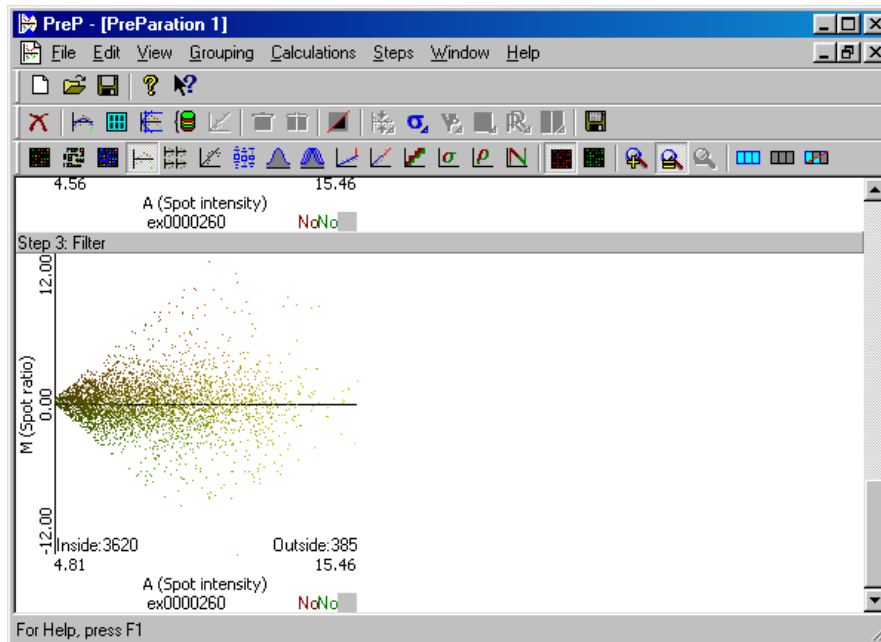
Now, click on the “setting threshold” icon:

A dialog box request the thresholds we wish to set and the values. Observe that a given threshold is active only when its corresponding *checkbox* is active.



The AM graph is displayed and the filtering icon is now active.

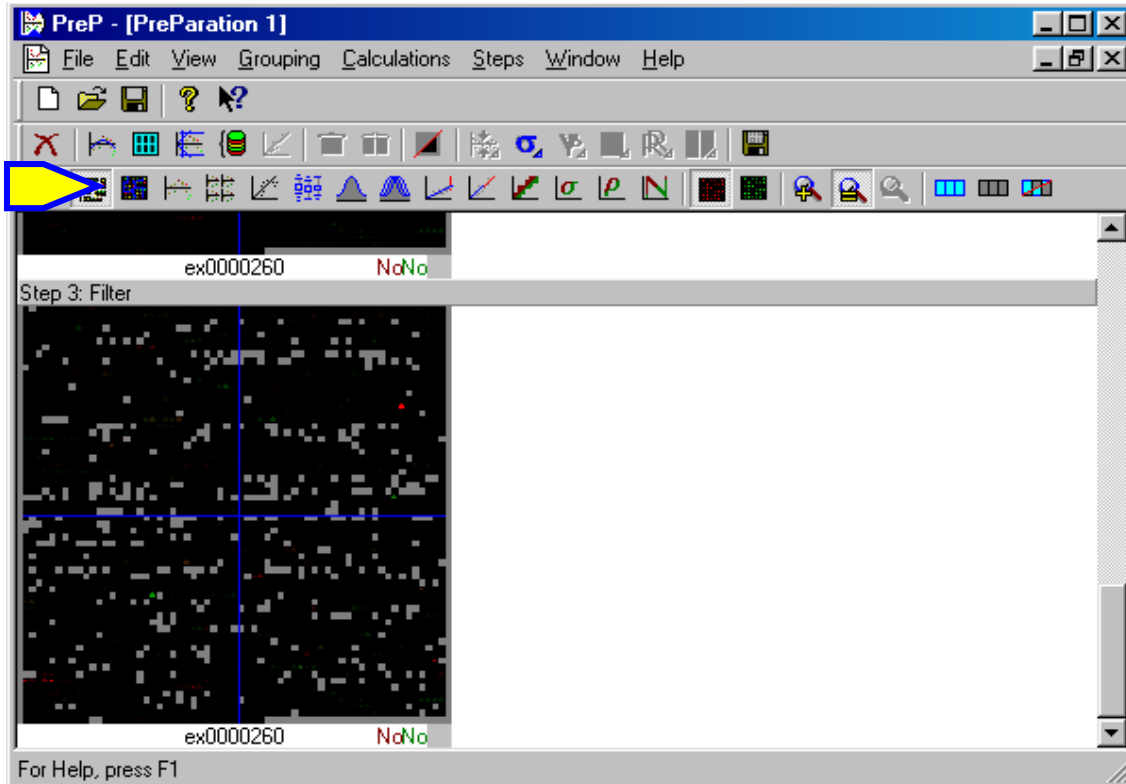
5.2.- Filtering



Once the thresholds have been set, PreP is ready to apply the filtering procedure. By clicking the icon, PreP will proceed to remove all spots out the thresholds scope. As can be deduced at this point, a new state is produced.



Using the “coherent view” of the slide the removed spots will appear as a grey spot..



6.- Closing words

There are other methods and procedures implemented in PreP, but they are described in the user manual where the reader will also be able to find extended information about the procedures mentioned on these pages.