

PreP+07

Gene-Expression Data Pre-Processing Tool

This document is an extension for the original User-Manual on previous versions.

The original version "PreP: gene expression data pre-processing" was developed by: *Jorge García de la Nava¹, Sacha van Hijum² and Oswaldo Trelles¹*; "PreP: gene expression data pre-processing"; *Bioinformatics* 2003 Nov 22; 19 (17): 2328-2329
¹Computer Architecture Department, Universidad de Málaga, 29080, Málaga, Spain; ²University of Groningen, Molecular Genetics, The Netherlands.

PreP+07(*) is an exhaustively enhanced version of PreP with new options that improve performance. This work has been carried out by Victoria Martín-Requena and Antonio Muñoz-Mérida with close support of Dr. G. Claros.

Victoria Martín-Requena, Antonio Muñoz-Mérida, Gonzalo Claros and Oswaldo Trelles; "PreP+07: an integrated software platform to improve gene expression data quality" (MS in preparation)

This document describes the new options, thus, original description in User-Manual and tutorials remains valid until this document modify the behaviour.

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

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

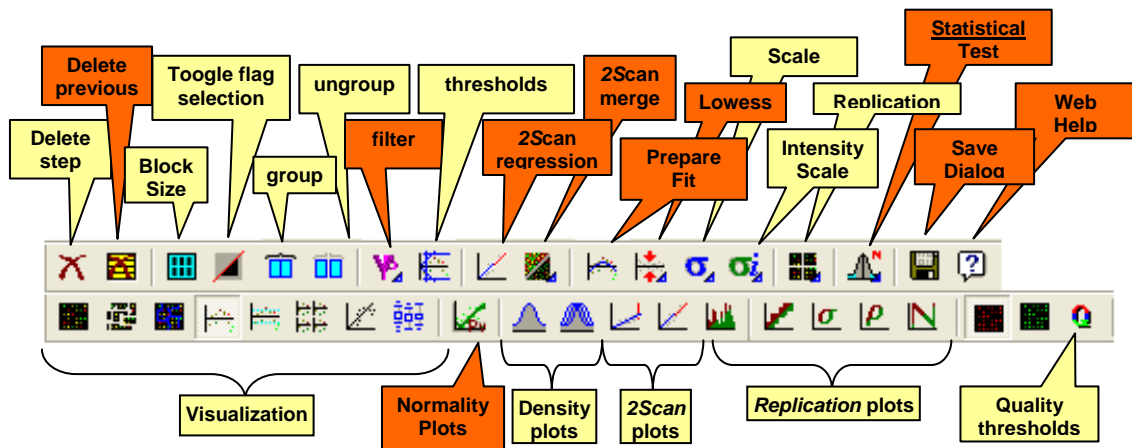
Bioinformatics and Information Technologies Laboratory, University of Málaga
<http://www.bitlab-es.com>



New functionality in PreP+07

1 Introduction

All the new operations are available in the toolboxes and the dialogs, a reorientation of the toolbars has been also done. New features in PreP07+ version are marked in orange.




- **Delete previous:** delete all steps except last.
- **New filtering methods.**
- **2Scan regression/2Scan merge:** first and second step to apply 2Scan method.
- **Prepare Fit/Apply lowess or S.Lowess:** new options to apply lowess or supervised lowess method.
- **Statistical Test:** New statistical test and reorganisation of the dialog.
- **Save dialog:** new save dialog with multiple functionality.
- **Normality plots:** Probability- probability plot, Quantile-Quantile plot or probability normal plot to asses the normality of the slides.
- **Web help:** a link to PreP's web help.

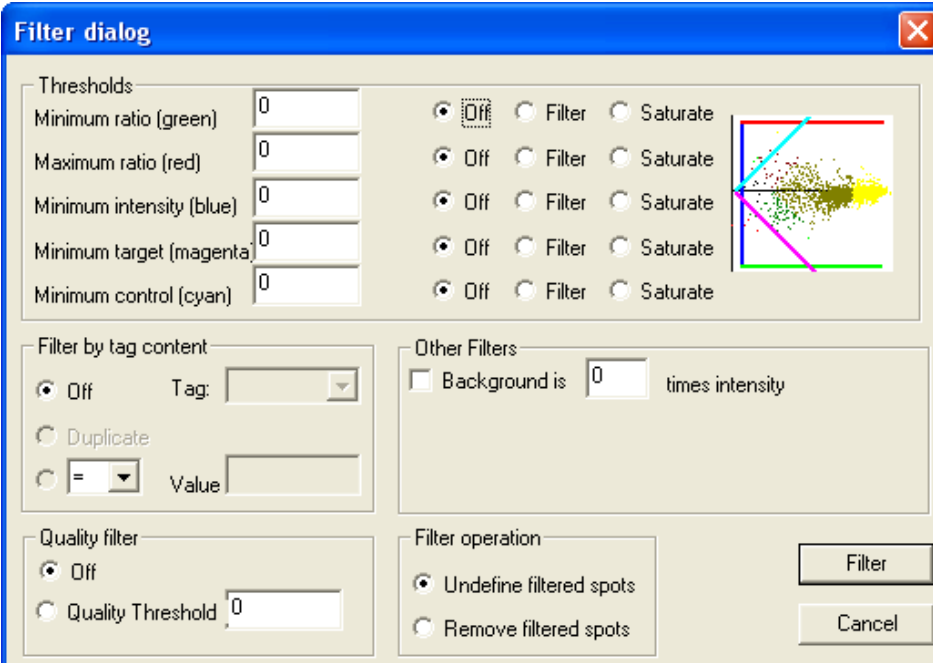
2 New filtering option

2.1 Description

In this new version it is possible to remove spots with a value “>, <, = or containing” a concrete value. Another new feature is to remove those points whose background is n times the intensity signal. Spots can be left undefined or removed. Old filters are still available.

2.2 Operation

Filtering is a step option (creates a new state) and is accessible through the  icon. This action will pop up the following dialog box.



Filter dialog

Thresholds

Minimum ratio (green) Off Filter Saturate

Maximum ratio (red) Off Filter Saturate

Minimum intensity (blue) Off Filter Saturate

Minimum target (magenta) Off Filter Saturate

Minimum control (cyan) Off Filter Saturate

Filter by tag content

Off Tag:

Duplicate

= Value

Other Filters

Background is times intensity

Quality filter

Off

Quality Threshold

Filter operation

Undefine filtered spots

Remove filtered spots

The meaning of the different groups are:

- **Threshold:** These options are only available if a threshold have been placed in the active state, in this new version these thresholds can be placed in the same dialog. The ways to deal with spots outside thresholds are:
 - **Off:** Do nothing.
 - **Filter:** Make the spot undefined or remove it (see Filter operation below)
 - **Saturate:** Move the spot to the threshold.

- **Filter by tag content:** A check will be performed in the selected tag name (in current snapshot the “description” tag is selected).
 - **Off:** Do nothing.
 - **Duplicates:** This will filter out the spots with duplicate tag values.
 - **</<=/>/>=:** This will filter out the spots whose tag contains the value that is greater/greater than/less/less than the string or number placed in the nearby edit control.
 - **has:** This will filter out the spots whose tag contains the string placed in the nearby edit control.

- **Quality filter:** It is just a very simple threshold filter.
 - **Filter by quality:** It activates or deactivates this threshold filter.
 - **Quality threshold:** The threshold value for quality. Spots with less quality will be removed.

- **Filter operation:** What to do when a spot is to be filtered.
 - **Undefine:** Leave the spot, but make all its values undefined. Since no spot is being removed, the index of them is kept. When saving, it is possible to remove these empty spots.
 - **Remove:** Take the spot off immediately.

- **Other filters**
 - **By background:** This filter will remove those spots whose background is n times the intensity, the n value must be configured in the text box.

3 Delete previous step

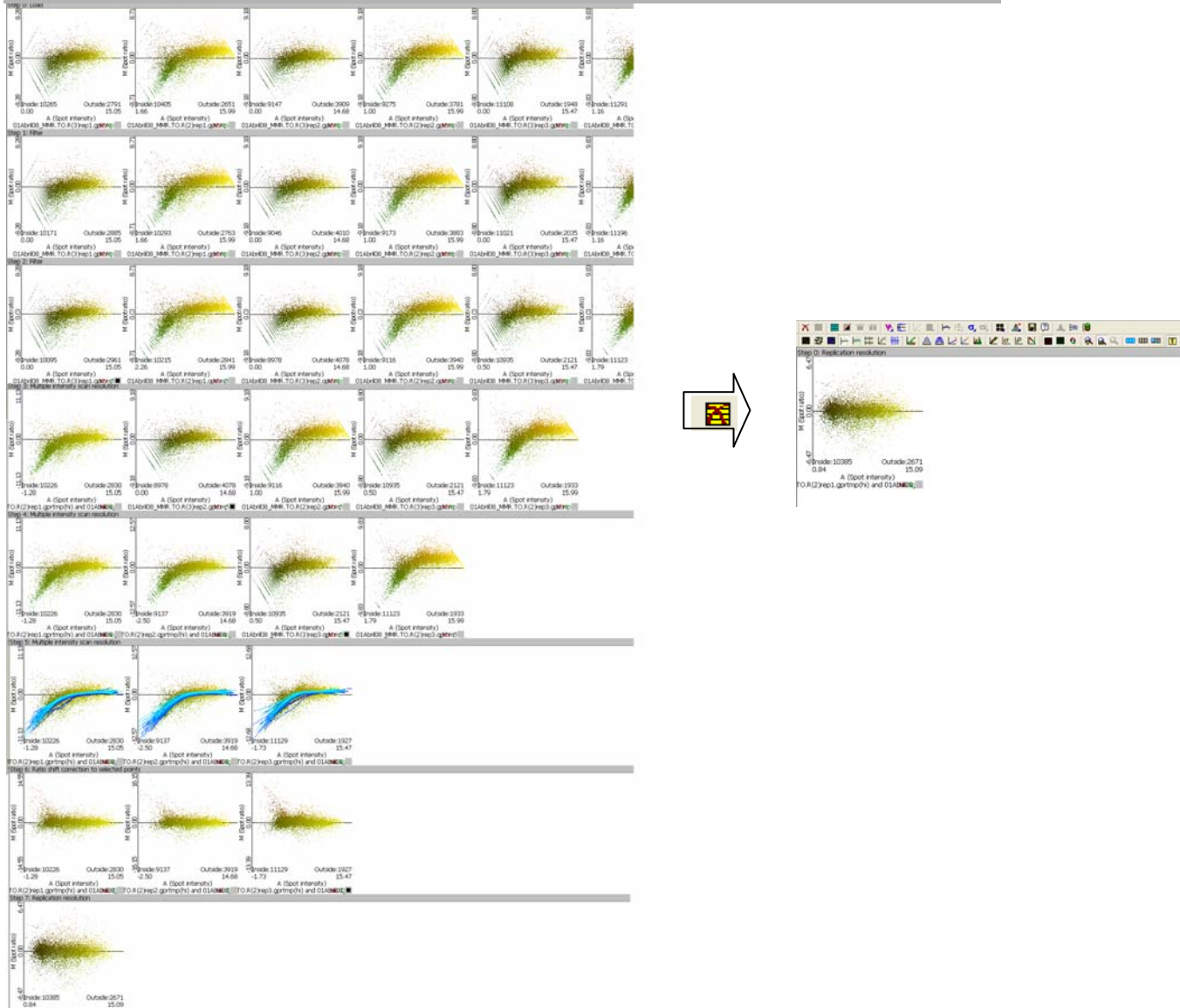
3.1 Description

Each operation made in PreP (not visualization) produces a new step, all steps are kept in a stack and visualized in PreP.

Sometimes this stack is too large and the researcher is not interested in the previous steps so this option allows the researcher to delete all the steps except the last one.

3.2 Operation

4



4 New replication method

4.1 Description

Dye-swap, interslide and intraslide replication is merged. Now all the spots with the same criterion inside a group are treated as replicates. These criteria are:

1. Spots on the same row in different slides of a group.
2. Spots with the same tag value inside a group (for instance, the same description)
3. Spots at the same place inside a group.

Several methods for solving replication are available:

1. Average (better if the number of replications is low)
2. Robust average (between average and median)
3. Median (better if the number of replications is high)

And each method can be used on additive or multiplicative (logarithmic) measures.

Also, a minimum number of replication may be required for a replication to be solved. The result is left undefined when the threshold is not reached.


Quality can be calculated in different ways:

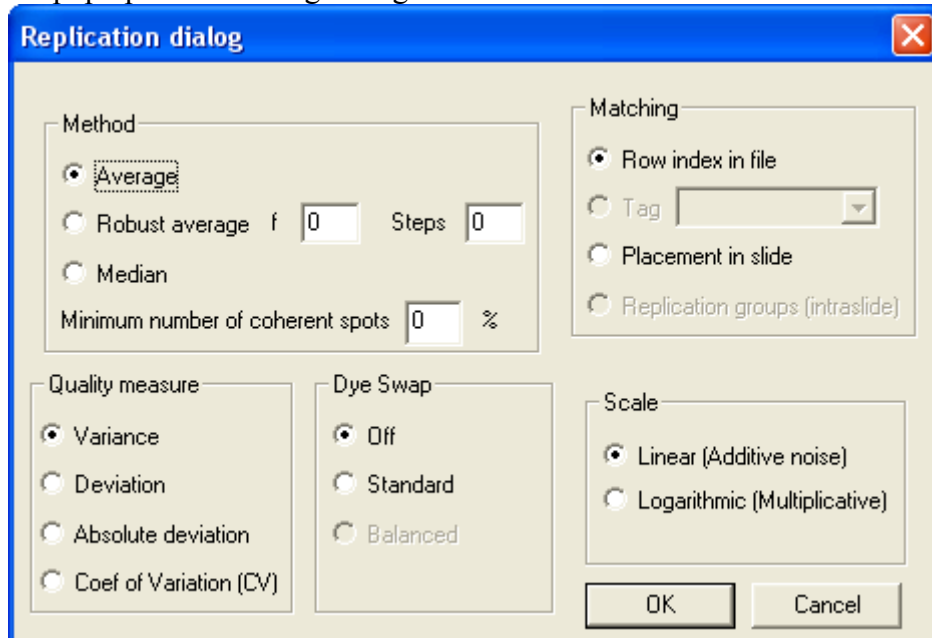
1. Negated variance (The more variance, the less quality)
2. Negated deviation (The more deviation, the less quality)
3. Negated absolute deviation (The more absolute deviation, the less quality)
4. Negated CV (The more CV, the less quality)

When a slide in a group is marked, that slide is though to be dye-swapped. There are two ways to handle a dye-swapped slide:

1. Ignore the dye-swap (off)
2. Swap the dyes (standard)


4.2 Operation

Replication is a step option (creates a new state) and is accessible thru the  icon. This action will pop up the following dialog box.



The meaning of the different groups are:

- **Matching:** How to know that two spots are replicates. First, slides which are not grouped are not to be used as replicates. Then, spot have to be matched. There are three ways of doing this:
 - **Row index in file:** Two or more spots are replicates if they are in the same row (in different slides).
 - **Tag:** Two or more spots are replicates if they have the same value in the selected tag (in different or the same slide).
 - **Placement in slide:** Two or more spots are replicates if they are in the same placement (in different slides).
- **Method:** When we know the replicated spots, they have to be solved into a single value. A minimum number of coherent spots (non-empty, non-undefined, that have brighter signal than background, etc.) may be required by **Minimum number of coherent spots** using a percentage of the total. Different methods for the solving can be used:
 - **Average:** Just perform the average.
 - **Robust average:** Use the average, then take the fraction f of closest spots and repeat **Steps** steps.
 - **Median:** Use the median (only when the number of replicates is high).

- **Scale:** The above resolution can be performed in linear or logarithmic space. The former for removing additive noise and the latter for multiplicative noise.
 - **Linear (additive noise):** Perform the solving using the intensity values.
 - **Logarithmic (multiplicative):** Perform the solving using the logarithmic values. *Better when you are using ratio logarithms.*
- **Quality measure:** Since we have a collection of replicated values, a statistical method can be applied for measuring the quality of that collection. Several classical methods are available:
 - **Variance:** Quality will be calculated as the negated addition of the variances of the replicated values for both channels.
 - **Deviation:** Quality will be calculated as the negated addition of the standard deviations of the replicated values for both channels.
 - **Absolute deviation:** Quality will be calculated as the negated addition of the absolute deviations of the replicated values for both channels.
 - **Coefficient of Variation (CV):** Quality will be calculated as the negated addition of the coefficient of variation of the replicated values for both channels.
- **Dye Swap:** When facing dye-swapped slides (marked in PreP by the bottom right squared  icon of each visualization), two behaviours are provided:
 - **Off:** Ignore dye-swap and use all the slides equivalently.
 - **Standard:** Swap the dye-swapped slides so all the same measures are in the same channel.

5 2Scan


5.1 Description

The devices used for measuring intensities are neither perfect nor unlimited. Saturation and quantization appear when scanning, and can barely be removed. Saturation appears due to the finite nature of the devices that have a maximum response value. When a measure exceeds this maximum, the device can only read its maximum and usually with distortion.

Quantization is due to finite word length in digital devices (e.g. scanners). A proprietary technique for improving the quality of scanned data (Garcia de la Nava, et al. 2004) is implemented in PreP+07, including the corresponding visualization as ‘intensity-intensity’ plot. It shows the spots according to the intensity measured in a low-sensitivity scan and a high-sensitivity scan for both the green and red channels. This plot can also be used for comparison of two replicated slides or scans. If the replication is properly done, the spots must show a linear relation; otherwise, when the scans have different calibrations, the data will follow a non-linear curve due to saturation or calibration effects .).

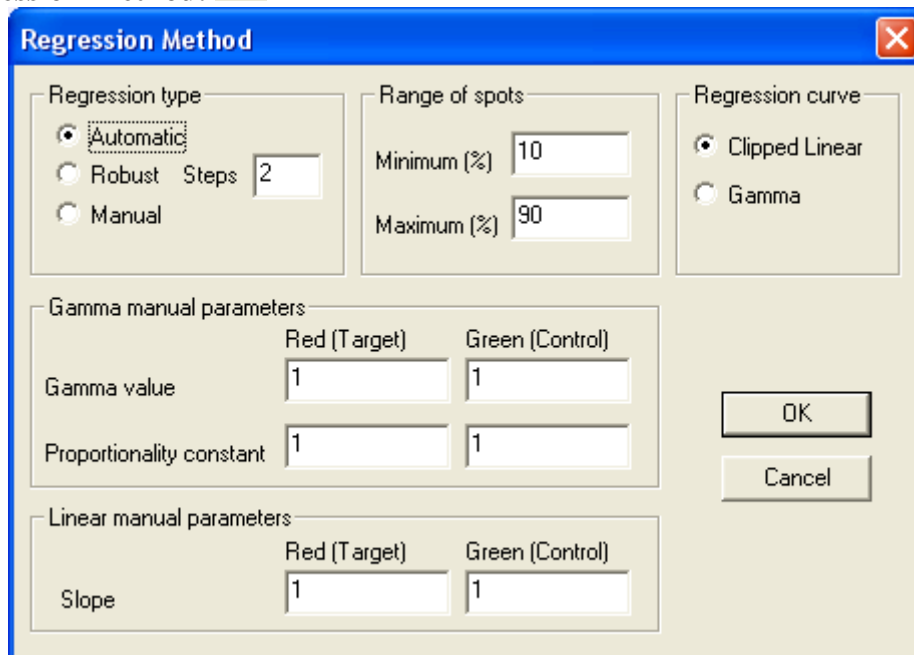
5.2 Operation

When different (intensity level) slide data are available, proceed by solving the double-scan. Let’s suppose we have n_s (non-saturated) data, and s_1 (saturated data):

1. Group non saturated and saturated slide: 

2. Mark non saturated as low intensity: 

3. Regression Method: 



The image shows a dialog box titled "Regression Method" with a close button (X) in the top right corner. The dialog is divided into several sections:

- Regression type:** Contains three radio buttons: "Automatic" (selected), "Robust Steps" (with a text box containing "2"), and "Manual".
- Range of spots:** Contains two text boxes: "Minimum (%)" with "10" and "Maximum (%)" with "90".
- Regression curve:** Contains two radio buttons: "Clipped Linear" (selected) and "Gamma".
- Gamma manual parameters:** A section with two columns: "Red (Target)" and "Green (Control)". It contains two rows of text boxes: "Gamma value" and "Proportionality constant", both with "1" in the boxes.
- Linear manual parameters:** A section with two columns: "Red (Target)" and "Green (Control)". It contains one row of text boxes: "Slope", both with "1" in the boxes.

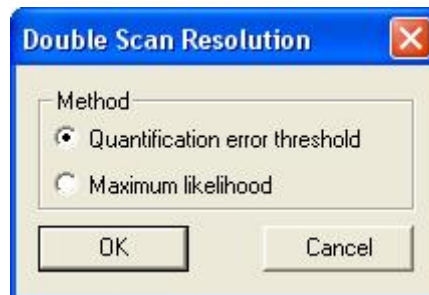
At the bottom right of the dialog, there are two buttons: "OK" and "Cancel".

- **Regression type:** The 2scan model is only useful in practical terms if the parameters can be determined or, at least, estimated. This can be achieved by:

- **Automatic:** a statistical regression.
- **Robust:** in view of the fact that outliers should be discarded, a robust regression is recommended.
- **Manual:** Gamma or Linear parameters are set by the user in the text boxes.

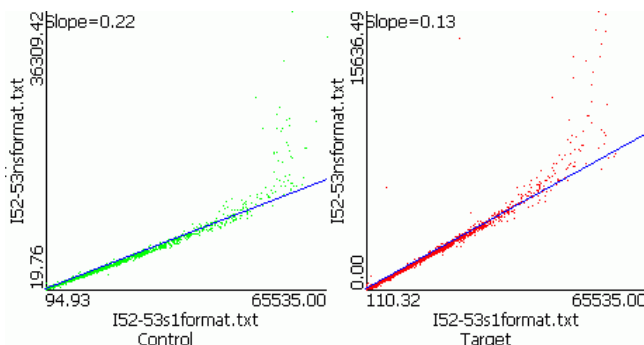
- **Regression curve:** The regression curve depends on the scanner settings, when a device is measuring a signal that is too intense, it becomes saturated and reports the maximum value it is able, even though the actual signal is higher. This is the ideal behaviour of saturation which is called 'clipping'. Clipping appears in some devices such as analog-to-digital converters (digitizers). Commonly the transition to saturation is more gradual and depending on its characteristics and the mathematical curve it follows, the saturation is said to be sigmoid, logarithmic, gamma, etc.

3. Multi-scan merge:



These two methods are described in the 2Scan paper available at <http://www.bitlab-es.com/prep>.

- **Visualization is also available through the icons:** 



6 Supervised Lowess

6.1 Description

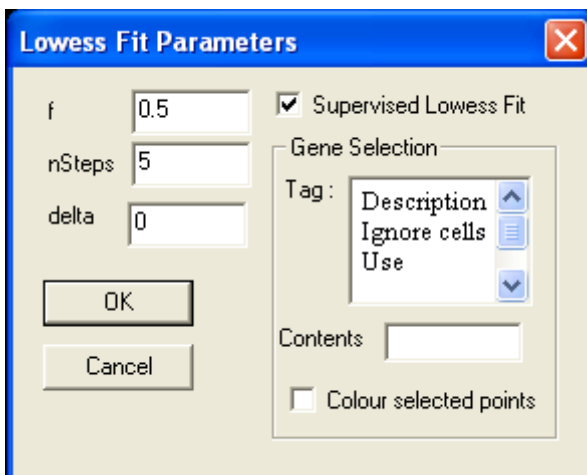
The supervised Lowess (SL) normalization method only uses genes that are conserved in both samples hybridized for normalization. In a first step the SL method perform Lowess normalization over the LHG subset of genes, computing the initial LogRatios (i.e. R_i ($i = 1 \dots N$)), followed by Lowess normalization, generating a set of corrected ratios R_{c_i} ($i = 1 \dots n$, $n < N$) and correction factors for the subset of conserved genes used: $\alpha_i = R_{c_i} - R_i$. Subsequently, the Lowess correction factors belonging to the subset of conserved genes (α_i) are extrapolated to determine the correction factors β_j ($j = n+1 \dots N$) for the remaining genes. The correction factors are then used to adjust the log-ratios of the remaining genes.

SL is one of the new methods included in the current PreP+07 version. Advantages of SL when data follow a non-normal distribution due to differences in gene sequence identity are demonstrated in (van Hihum et al. 2008), suggesting that it is appropriate for any microbial aCGH comparison. In any case, the supervised lowess assess a normalizing estimate using a sub-set of genes (sharing strong sequence similarity) and then uses this estimation to remove the error in the rest of genes. This procedure has been successfully applied to spiked-in dual dye DNA microarray data.

6.2 Operation

An extended manual of this feature is available at <http://www.bitlab-es.com/prep>.

a. Prepare fit

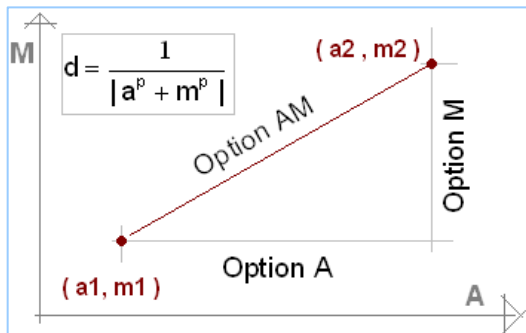
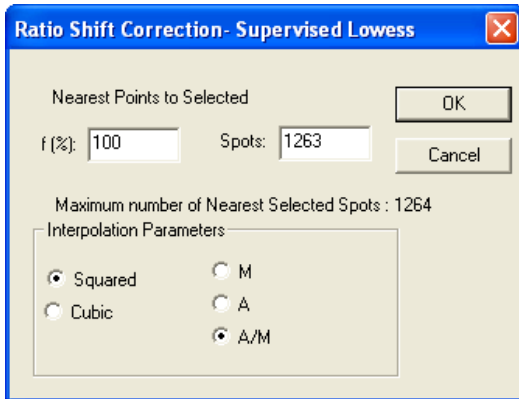


Supervised Lowess has been implemented using the Lowess icons, applying Supervised Lowess or Lowess method is an option in the dialog box. This option is only available when tags are loaded.

To apply S.L. method select the Supervised-Lowess option and the “Gene Selection zone” will become active. Specify the filter-column. Additionally, you must specify the “selection” value. Genes will be used in the Supervised-Lowess if they contains that value.

b. Ratio shift correction

PreP+07 is able to recognize whether a Supervised fitting was performed (or not) as previous step.



The interpolated ratio shift for ‘not selected genes’ is computed based on the nearest ‘Selected Genes’ to each spot. PreP+07 need the percentage of the Selected Spots or the number of Spots (both values are synchronized) in such a way that if we change the percentage (f%) the number of Spots in the other field is automatically updated and vice-versa.

The interpolation parameters are used to control the Interpolation method and distance. On the left a distance scheme is shown. The default distance is Cartesian, but can also be based only y the A or M values. Square or cubic exponent can be used to compute distance.

7 Statistical Test

7.1 Description

PreP+07 allow performing the statistical tests through two methods:

- Global estimation: calculates the average and std. deviation using all spots (classical approach).
- Using local variance (like SNOMAD or other differential expression software).

A reorganization of the old statistical dialog has also been performed; a single step is now needed to calculate z-scores/p-values.

Thus, the first step is always estimating this local variance. There are several ways to calculate the local variance:

1. Windowed estimation of standard deviation (The classical approach)
2. Lowess estimation of absolute deviation (Robust approach)
3. Lowess estimation of standard deviation (Robust approach)

Sometimes is interesting to distinguish between positive and negative values. This option is also included.


Once the local variance has been estimated, a statistical test can be used. There are several methods available at this step also:

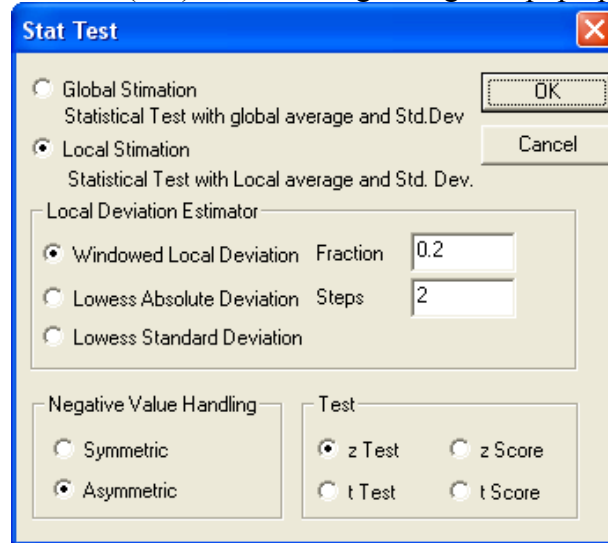
1. z-Test (default option) Use when the number of spots is high.
2. t-Test (only if using windowed estimation of standard deviation)
3. z-Score. Calculate p-Value, but keep z-Score as quality (using variance estimator).
4. t-Score. . Calculate p-Value, but keep tz-Score as quality (using variance estimator).

For allowing a descriptive visualization of the p-value a quality MA graph have been integrated. This graph has two modes of drawing:

- Continuous colouring (each quality value has a colour)
- Threshold colouring (the user can select two threshold levels for quality)

7.2 Operation

Using the statistical test icon () the following dialog will pop up:



Two options are available the global estimation and local estimation, if local estimation is selected, the groups of options “Local deviation estimator” and “Negative Value Handling” will become active, otherwise these options won’t be available.

- Local Estimation

- **Local Deviation Estimator:** Is the method to be used for the local deviation estimator. There are three
 - **Windowed Local Deviation:** It takes a **Fraction** of spots near the spot whose deviation is to be found, and it use that local spots for the estimation.
 - **Lowess Absolute Deviation:** It uses a lowess curve (given it **Fraction** and **Steps**) for absolute deviation fitting.
 - **Lowess Standard Deviation:** It uses a lowess curve (given it **Fraction** and **Steps**) for standard deviation fitting.
- **Negative Value Handling:** This selection shows what to do with the negative ratios.
 - **Symmetric:** Force the deviation to be the same for positive and negative values.
 - **Asymmetric:** Allow different deviations for positive and negative values.

Performing the statistical test is just a matter of selecting which the type of test in the Test group of options.

8 Probability Plots

8.1 Description

There are several ways to assess the normality hypothesis in a data set, visual inspection through plots is a useful method. These are graphical techniques for assessing whether or not a data set is approximately normally distributed. The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line.


Three different normality plots are well known:

- PP (probability probability): p-values vs. expected p-values.
- QQ (quantile-quantile): z-scores vs. Expected z-scores.
- PN (probability normal): logratios vs. expected p-values.

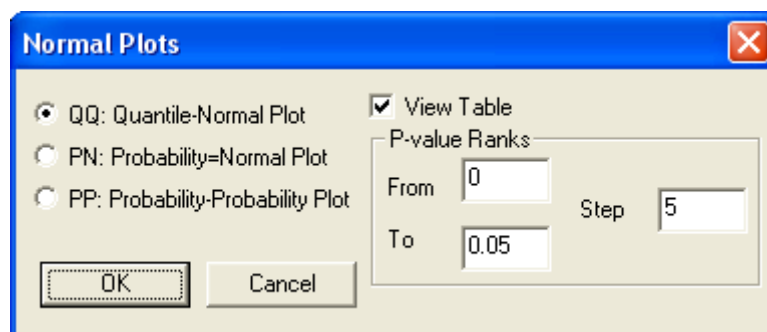
These plots represent data pairs, where each pair is conformed by an observed value in the data set and the value that must correspond if the data set fit exactly a normal distribution (expected value).

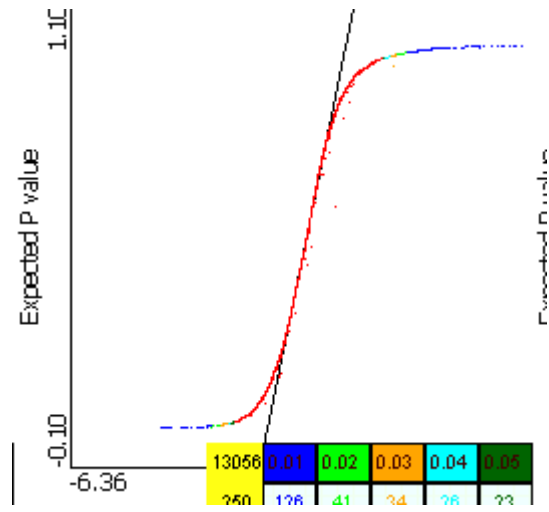
If the data set fit a normal distribution all spots must be in the 45° diagonal, the spots that outlie that diagonal line are the differential expressed ones.

8.2 Operation

To assess the normality of the slides just click on the normality graph icon  and select the plot.

An extra option is available for some plots, this option is to show a table with the number of spots that are in a p-value interval.





9 Tag parsing

9.1 Description

Some tags contain more than one field. For instance:

Plate 2 [IL1403_384_04], Well P22 [il1403_yndE]

This tag value contains: plate number, plate content, well position and well content.

9.2 Operation

When using the tag functionality, the character @ is used for opening and closing a matching string. The first character of the tag must be @ for matching to occur. For instance:

@[@Plate Content@]@

This is a matching pattern that will look for a “[”, then it will create a column called “Plate Content” until a “]” is found.

Several matching characters can be placed and also several column names:

@[@Plate Content@],[@Well Content@]@

First, it will look for a “[”, then the plate content until a “]” is found. After that, a comma will be sought, and the next “[”. The well content is then read until another “]”.

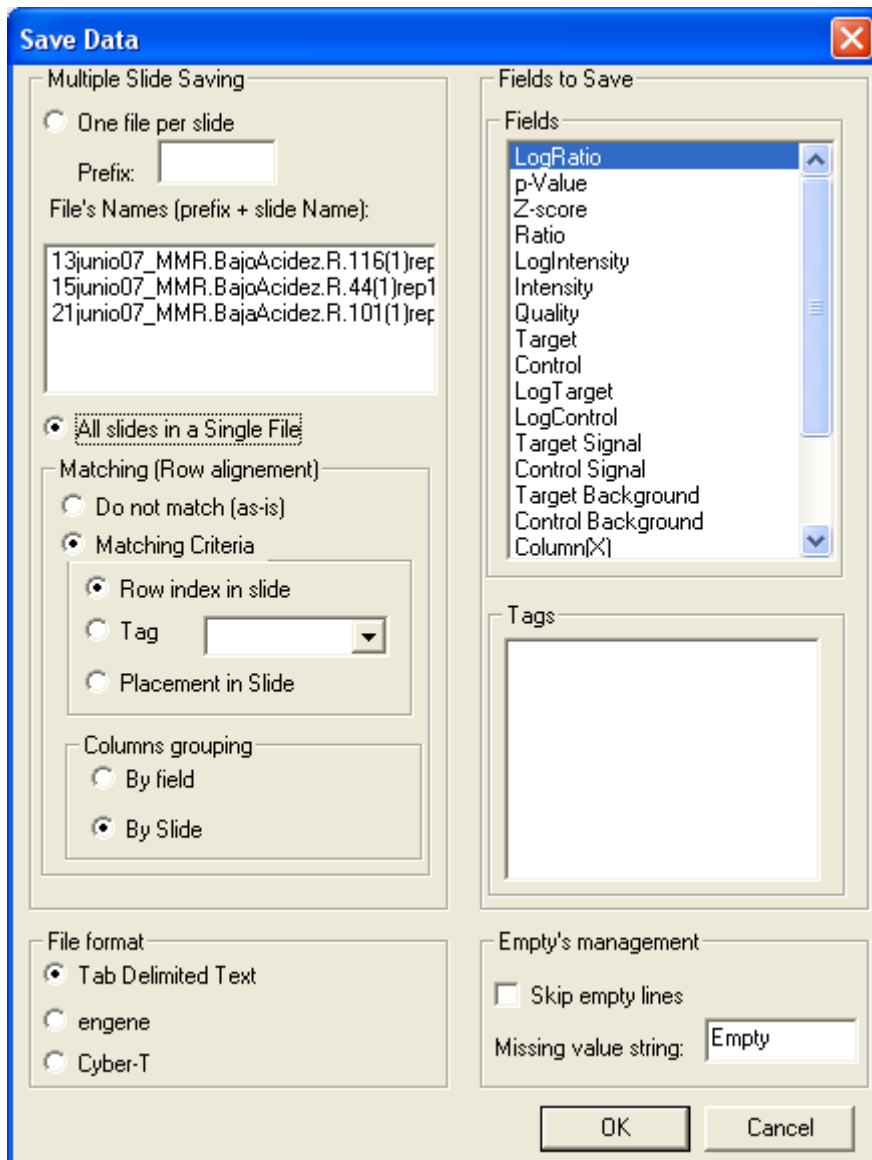
10 New saving data options

10.1 Description

Some small modifications have been built in the saving options for user convenience. Multiple or single file saving and the possibility of writing the p-value and the z-score directly at the output file.

Compatibility with CyberT and engine is granted via an special button, in the case of CyberT PreP will report an aid for using CyberT on the saved data.

10.2 Operation



The save dialog box has some new options. The additions are:

- **Multiple Slide saving**
 - **One file per slide:** use one file per slide loaded in PreP.
 - **All slide in a single file:**
 - **Matching (Row alignment):** Now you can save several slides and place in the same row spots with a common feature (row index, tag coincidence or placement). When matching tags, a new column with that tag will be automatically added.
 - **Columns grouping:** this option is needed to select the order of the columns in the output file, all columns of the same slide together or all column of the same variable (i.e. logratio) together.
- **Fields to save**
 - **p-Value/z-score:** These are new available fields that will save the p-value and/or z-score (if a statistical test was performed).
- **File format**
 - **Tab delimited text:** single row of column headers, data separated by tabs.
 - **CyberT:** it will show an aid for using CyberT and it will save the data in a CyberT compatible way.
 - **Engene:** Special file format compatible with engene software.

Like in previous versions, descriptions are to be selected if you want them included in the resulting output file.