Bootstrapped ROC (bROC) Plugin Manual

Overview

bROC plugin is used in the discovery of differentially expressed probes/genes in microarray and RNA-seq (next generation sequencing) experiments. It deploys in CLC Main Workbench and CLC Genomics Workbench.

The *bROC* compares expression between two biological states/endpoints (e.g., treatment and control samples, disease and normal, etc.). At least two experimental/biological replicates are required per state. The algorithm is especially useful for analysis of data sets with small number of replicates and large number of features/probes (thousands or, tens of thousands).

In general, both microarray and RNA-seq data should be normalized before the differential expression analysis is performed. *bROC* plugin includes normalization procedure optimized for the algorithm and the RNA-seq data. Thus, for RNA-seq data, a workflow consisting of *RNA-seq Analysis* tool and *bROC* provides complete analysis of differential expression, from raw reads to the list of expressed genes.

If needed, the input data are automatically log2 transformed before they are used in the analysis. For RNA-seq data, which contain null values, the typically used automatic transforms are unity shift and log2.

ROC (receiver operating characteristic) is a generally applicable, non-parametric procedure that provides insight into the discriminatory properties of data features for a binary classifier. However, the method is not efficient for gene expression experiments as they generally do not produce a sufficient number of samples. *bROC* overcomes this limitation by resampling (bootstrapping) the expression data to produce a large number of 'simulated' measurements that preserve the statistical properties of the original data. Thus, *bROC* can produce detailed curves of sensitivity (probability of true positive detection) vs. 1-specificity (probability of false positive detection) for all features of interest. The area under the curve (AUC) is the primary statistics used for detection of regulated features (probes, genes). For each probe/gene *bROC* produces the following statistics (discrimination scores):

- CONF = 2 AUC – 1 where AUC is area under the ROC curve. CONF (also known as Gini coefficient) is equal to twice the area between the ROC curve and the no-discrimination line. CONF = 1 (AUC = 1) indicates perfect separation of the expression measurements between two states and CONF = 0 (AUC = 0.5) indicates no separation (i.e., no differential expression).
    - CONF is recommended as the primary statistics in differential expression studies, with detection threshold typically set at 0.95 (probes/genes with CONF > 0.95 are considered to be differentially expressed).
- PD (probability of detection) balanced against PFA (probability of false alarm). This value is calculated at the intersection of ROC curve and diagonal of the ROC plot (PFA $\cong 1 - \mathrm{PD}$).

Other outputs include:

- Estimates of standard deviation for CONF and PD.
- Fold change used in the analysis. In most cases *bROC* automatically transforms the input data (see *Automatic transformations of input data*) and thus, this quantity may be different from fold change presented by CLC Workbench for the original data.

Optionally, the following plots are produced:

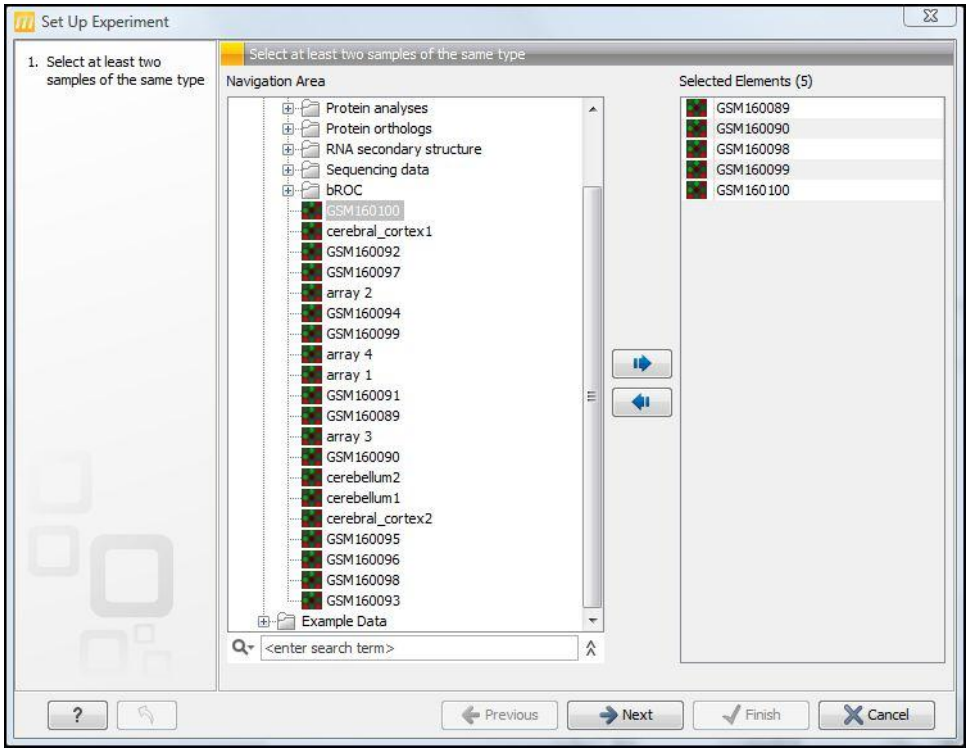
- MA plot (difference between average group values vs. overall average expression), showing the features declared as differentially expressed for a given CONF threshold.
- XY plot (average of one group expression values vs. the other group), with differentially expressed features.
- CONF vs. Fold Change ('Volcano' plot).

The *bROC* algorithm has been validated on microarray (Affymetrix GeneChip) and next generation sequencing (Illumina Genome Analyzer) data.

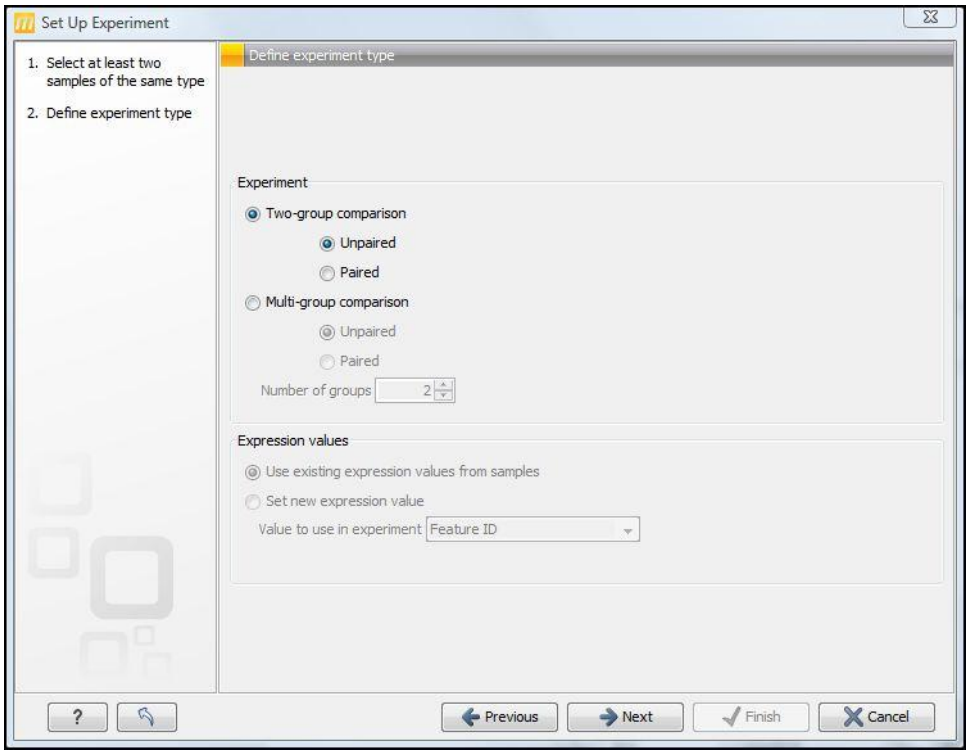
Experiment Setup

Main Workbench. In Toolbox select: Expression Analysis → Setup Experiment
Genomics Workbench. In Toolbox select: Transcriptomics Analysis → Setup Experiment

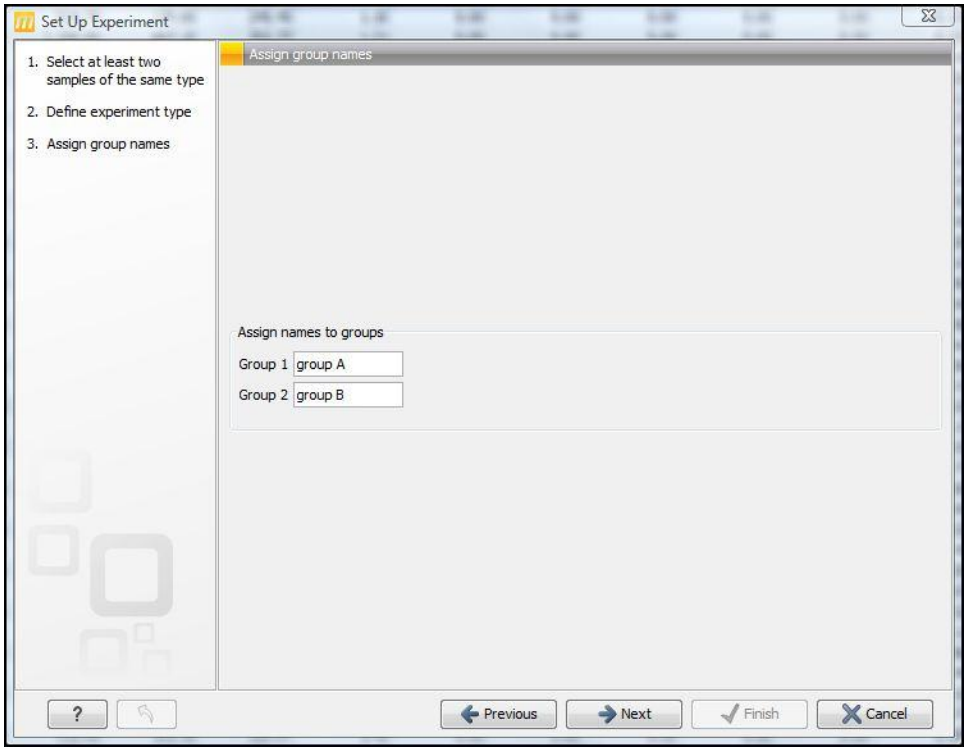
Step 1: Select at least four (at least two per group) samples of Expression data.



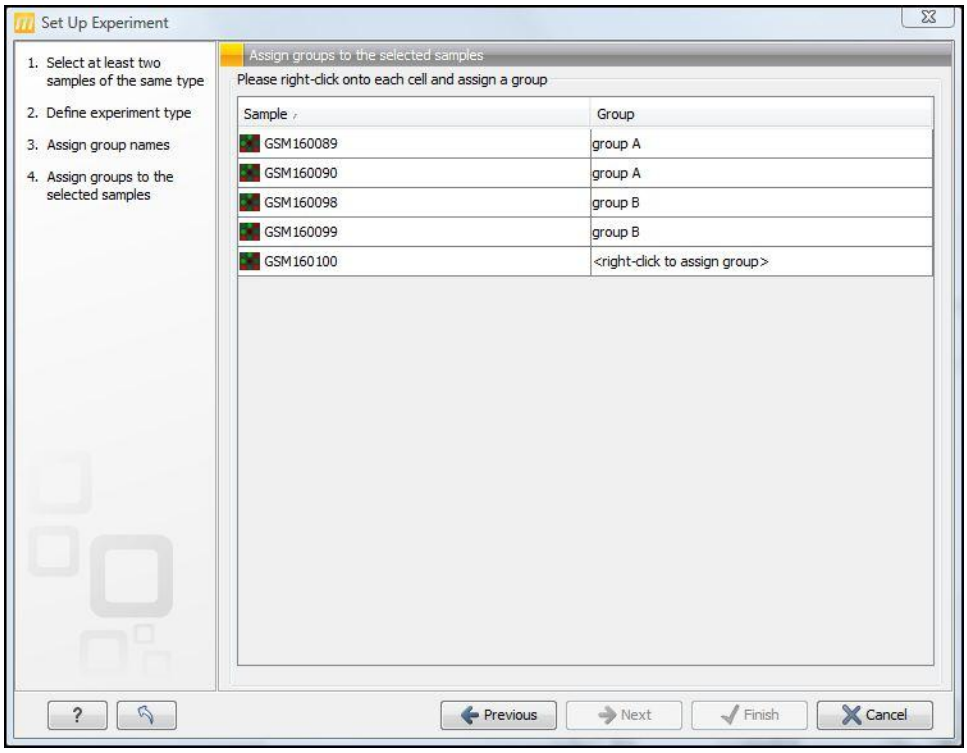
Step 2: Select 'Two-group comparison' and 'Unpaired'.



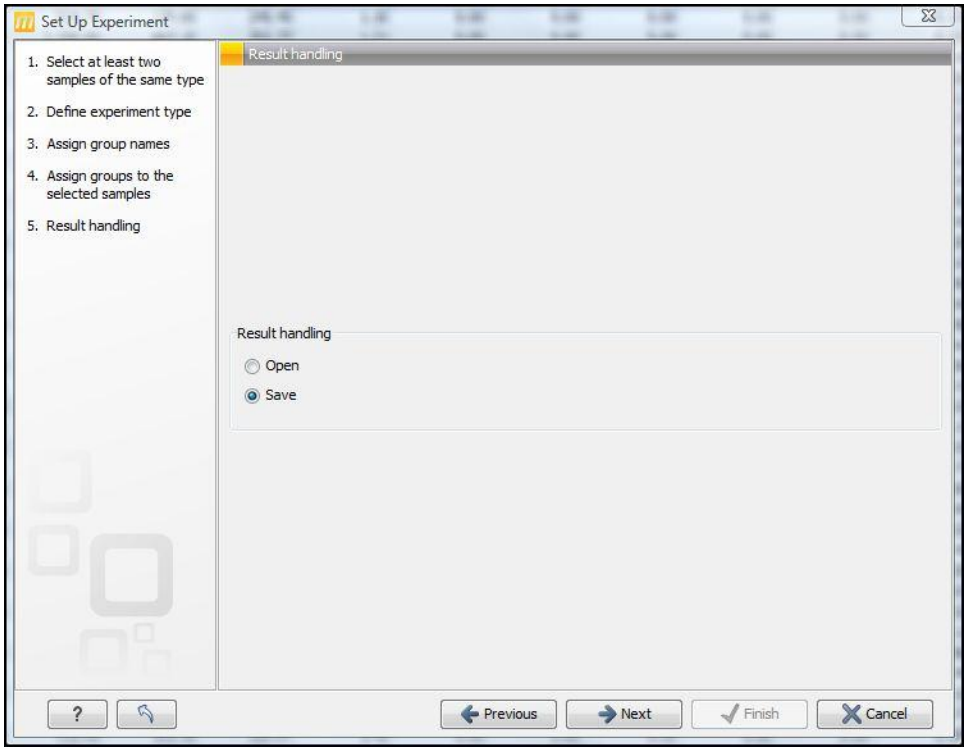
Step 3: Choose names for the two groups. (Here, we have group A *and* group B*)*



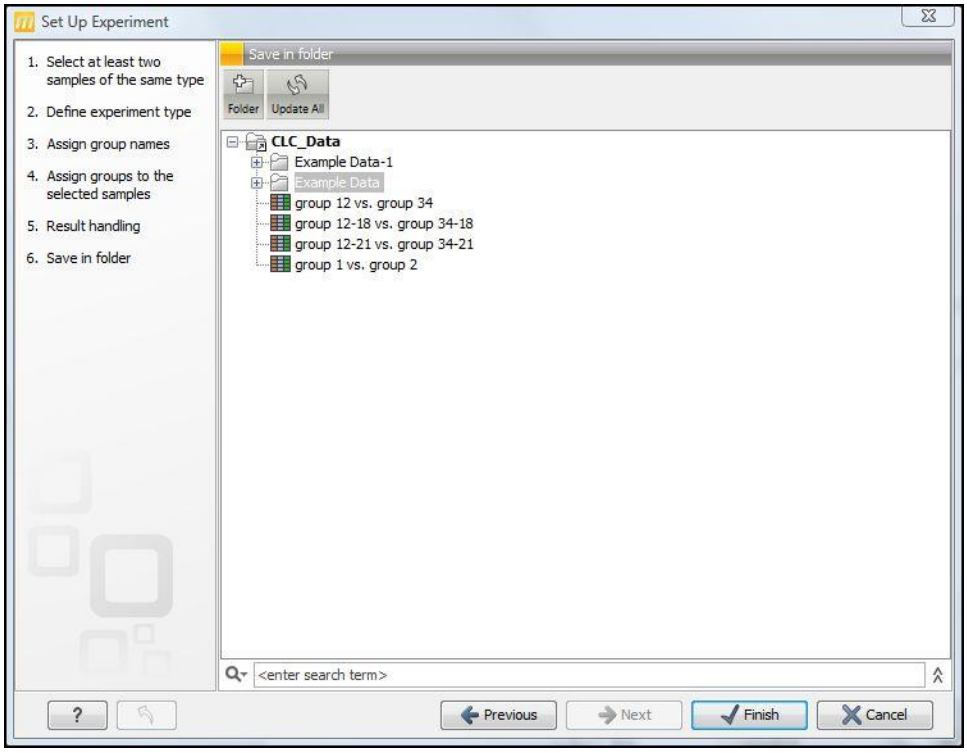
Step 4: Assign at least two samples to each group. Click Next.



Step 5: Select 'Open' or 'Save' for Result Handling (the experiment may be saved at this step or later).

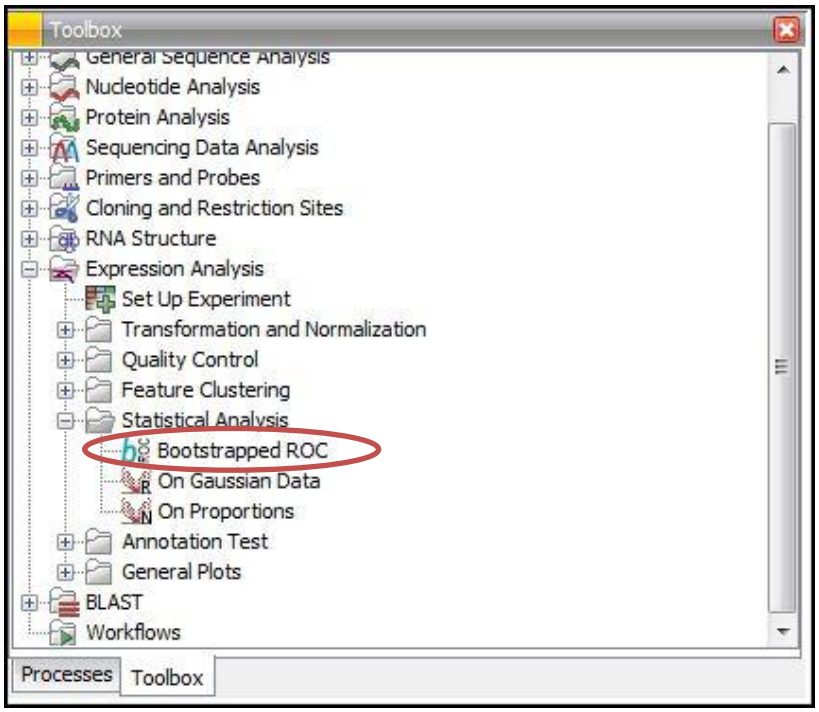


Step 6: For 'Save' select the folder where the experiment should be saved.

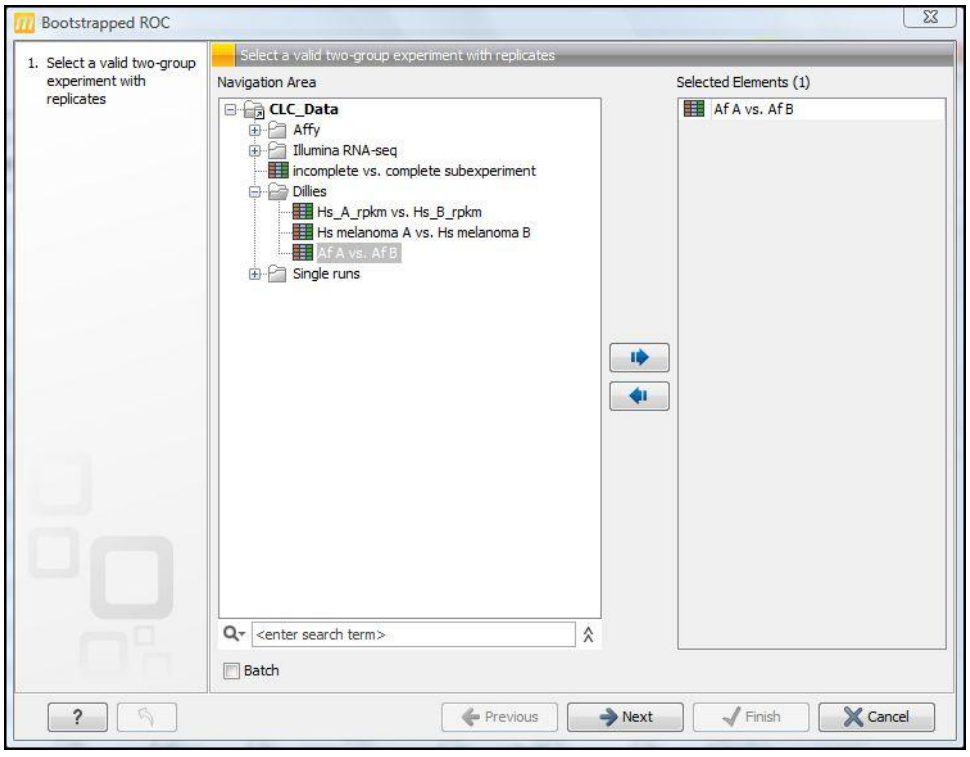


Running bROC

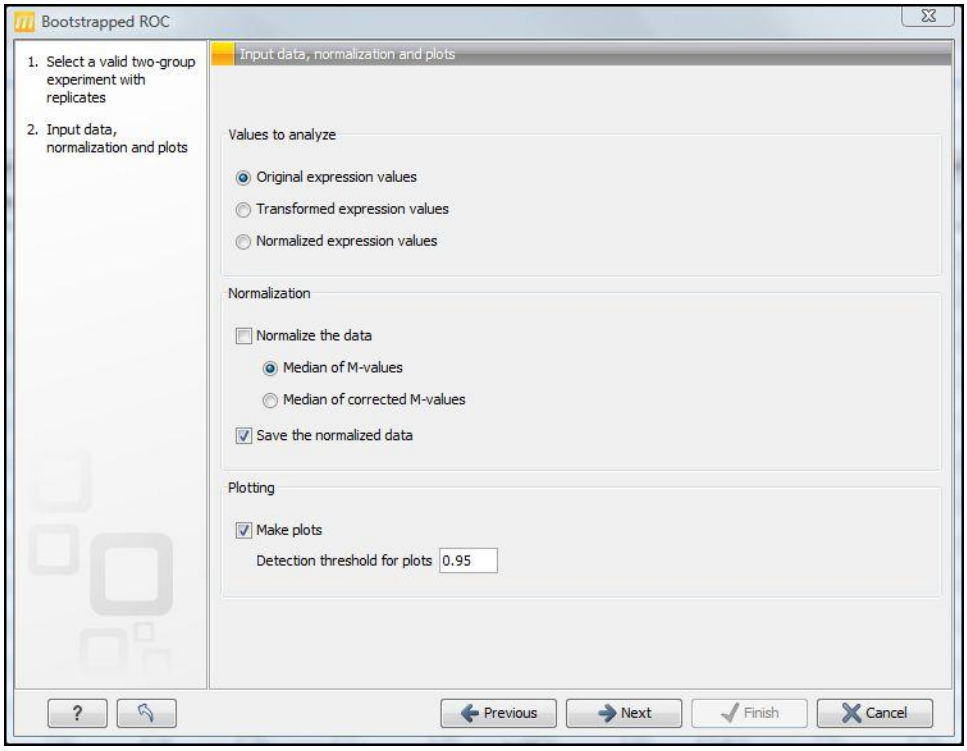
Main Workbench. In Toolbox select: Expression Analysis → Statistical Analysis → Bootstrapped ROC
Genomics Workbench. In Toolbox select: Transcriptomics Analysis → Statistical Analysis → Bootstrapped ROC



Step 1: Select a valid experiment (two groups with at least two samples/replicates per group).



Step 2: Input data, normalization options and plots



Values to analyze. Select from available data types. *Transformed Expression* data and *Normalized Expression* data are produced by the Workbench Transformation and Normalization tools, respectively, and thus, may not be available. This option is provided to allow the user some flexibility in data pre-processing. *Original Expression values* is the default.

Normalization. The inter-sample normalization method implemented within bROC is ***Median of M-values (MMV)*** – it employs the median of M-values used in the analysis. The raw RNA-seq count data are transformed (see *Automatic transformation of input data* below), and M-value for sample *n* and gene (feature) *g,* M_{ng}, is given as:

$$M_{ng} = \log_2 (E_{ng} + 1) - \log_2 (E_{0g} + 1) = e_{ng} - e_{0g},$$

where: E_{ng} is expression value (count number) for sample *n* and gene *g*, E_{0g} is the expression value for a reference sample, and e_{kg} are the respective transformed values. Median over all genes, for which expression is not null for at least one of the considered samples, produces the normalization factor, N_n for sample *n*. The first sample in the experiment table is used as the reference. The normalized expression values $<e_{ng}>$ are obtained as:

$$<e_{ng}> = e_{ng} - N_n.$$

The MMV normalization is intended primarily for RNA-seq data and has not been rigorously tested for microarray data. Nevertheless, at least in principle, this normalization approach, when followed by bROC analysis, should be applicable to all expression platforms, including RNA-seq and microarrays.

Save the normalized data is selected to save the normalized data (which are, typically, also log2-transformed) as *Normalized expression values.*

Plotting. Use *Make plots* option to produce and save plots that are helpful in visualization of bROC analysis results:

- MA and XY plots showing the features detected by bROC as differentially expressed. MA plot shows difference of expression between two groups versus group-averaged expression. XY plot depicts average expression in one group versus the other. The default detection threshold for the plots is 0.95 – the features with CONF ≥0.95 are declared as differentially expressed. The threshold value may be

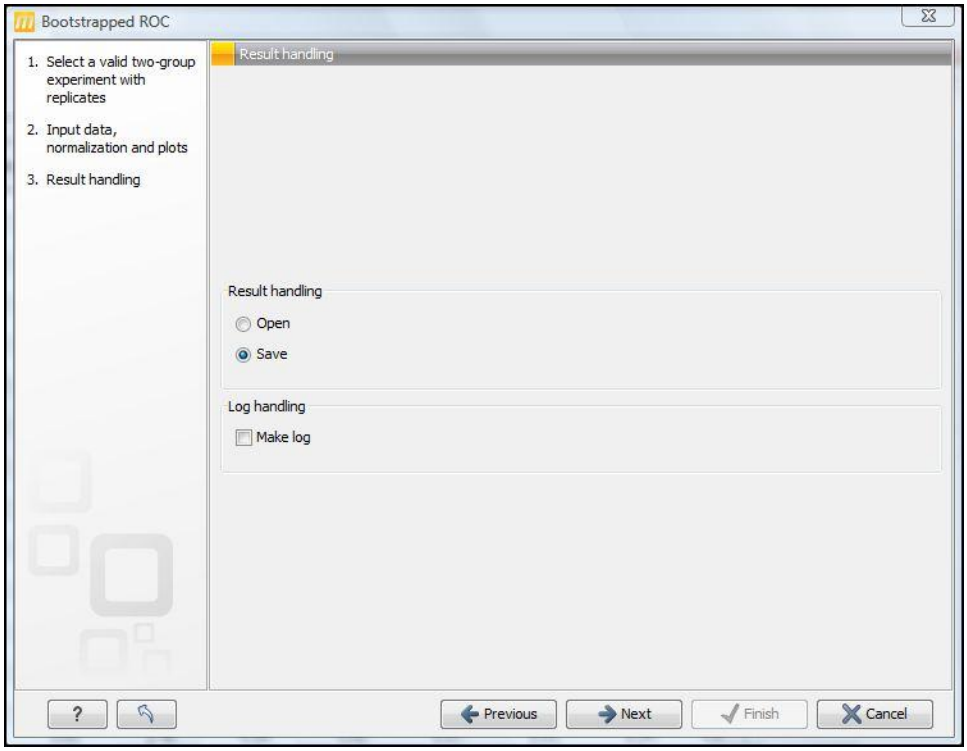
changed here and it changes the plot appearance, only – more features will be depicted as differentially expressed for smaller values of the threshold.

- The 'volcano' plot depicts CONF vs. Fold Change, the latter as calculated and used by the bROC algorithm, equivalent to the M-value for group-averaged data. Mouse pointer placed over a datapoint invokes display of the datapoint/feature name, and Fold Change and CONF values.

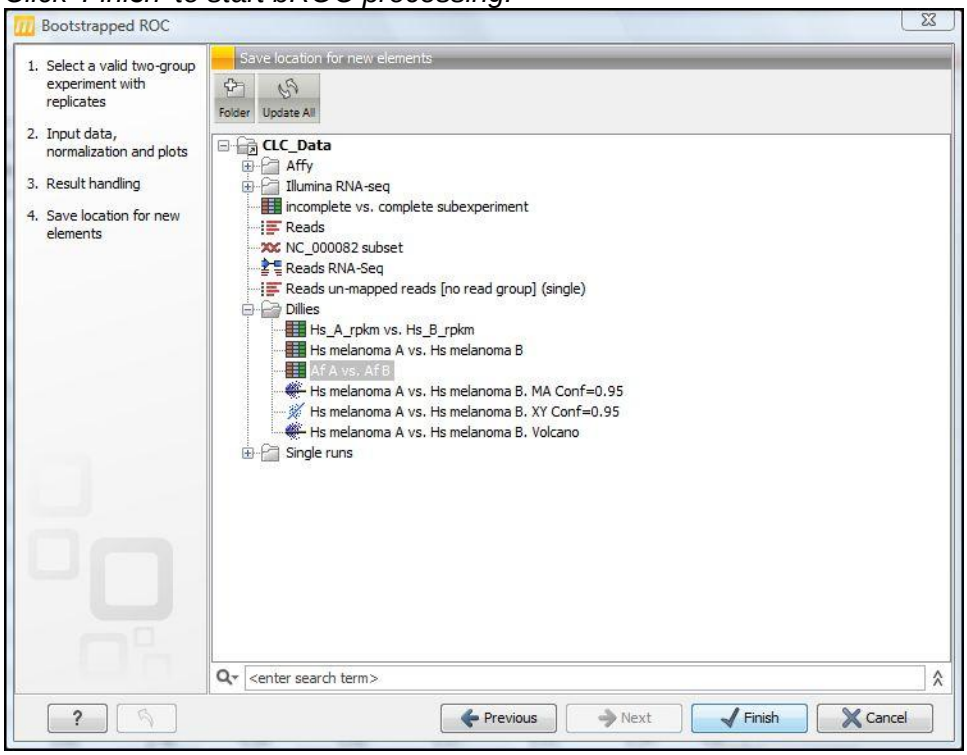
Plots are saved in the same folder as the experiment file - see Step 4.

NOTE: use *Plot Settings* to adjust axis' ranges, color and type of datapoint markers, and overall appearance of the plots.

Step 3: Select 'Open' or 'Save' in 'Result handling'. If you save the results, additional columns of bROC analysis will be added to your experiment ('Af A vs. Af B' in this example).



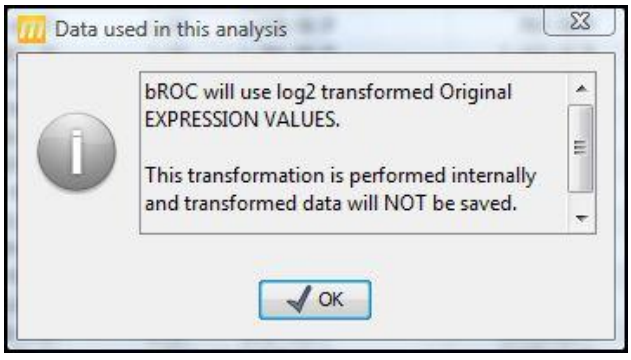
Step 4: Select the folder or file where results should be stored. Plots (see Step 2) will be stored in the same folder.
Click 'Finish' to start bROC processing.



Automatic transformations of input data

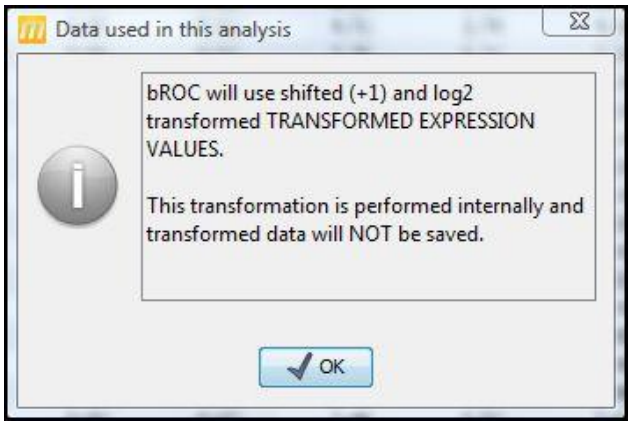
bROC requires log2-transformed data. The plugin checks the data values to infer if the log2-transformation has been performed.

- If not, log2-transform is applied and transformed data are used in the analysis. If the *Original expression values* are used (see above) the following information message is produced:



The fold change presented in the result table is Fold Change = $\log_2(I_A/I_B)$, where I_A and I_B are mean expression values measured for Group A and Group B, respectively.

- If log2-transformation is required but the input array includes null values (as is the case for RNA-seq data), shift transformation (+1) is applied first. For example:



Fold Change = $\log_2 [(I_A + 1) / (I_B + 1)]$.

- Otherwise, the selected input values are employed without any additional transformations (e.g., when *Transformed expression values* are used as input). Fold Change = $I_A - I_B$.

NOTE: Results of the internal transformations are NOT saved. Only the fold change calculated for transformed data is saved and presented in the result table.

Release note for Version 3

Important changes with respect to Version 2 include:

- The bROC algorithm was optimized for count (RNA-seq) data.
- Computation of normalization factors (for RNA-seq data) was included as an option.
- Data selection, normalization and plotting options were added – see *Running bROC, Step 2.*
- Automatic selection of input data was removed.

BioFormatix
12396 World Trade Drive, Suite 315
San Diego, CA 92128
USA
info@bioformatix.com
www.bioformatix.com