Genstat®

Analysis of microarray data

www.vsni.co.uk

VSNi

Analysis of Microarray Data
(22nd Edition)

by David Baird.

Genstat is developed by VSN International Ltd, in collaboration with practising statisticians at Rothamsted and other organisations in Britain, Australia, New Zealand and The Netherlands.

Published by: VSN International, 2 Amberside, Wood Lane,
Hemel Hempstead, Hertfordshire HP2 4TP, UK
E-mail: info@genstat.co.uk
Website: http://www.genstat.co.uk/

First published 2006, for GenStat *for Windows* 9th Edition
This edition published 2022, for Genstat *for Windows* 22nd Edition

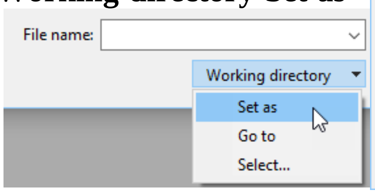
Contents

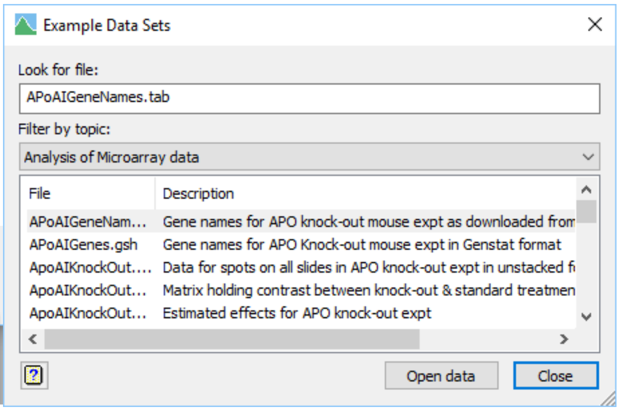
Microarray example files

The microarray data files must be downloaded separately from https://kb.vsni.co.uk/wp-content/uploads/Microarrays.zip. These should then be unzipped to the folder C:\Program Files\Gen22Ed \Data (this will create a Microarrays folder under Data). If you do not have rights to unzip files to that directory, then they can be placed in any directory, but will not be found in the File | Open Example Data Sets menu. If you are unsure of how to unzip the files, then opening the Microarrays.zip file with File | Open will let you select a file from the zip file. The Microarrays folder should contain the following files:

File	**Description**
APoAIGeneNames.tab	Gene names for APO knock-out mouse expr as downloaded from web
APoAIGenes.gsh	Gene names for APO Knock-out mouse expt in Genstat format
ApoAIKnockOut.gsh	Data for spots on all slides in APO knockout expt in unstacked format
ApoAIKnockOutContrast.gsh	Matrix holding contrast between knock-out & standard treatments
ApoAIKnockOutEffects.gsh	Estimated effects for APO knock-out expt
ApoAIKnockOutSlides.gsh	Description of 16 slides used in APO knock-out expt
ApoAIKnockOutStacked.gsh	All data in APO knock-out expt
APoAISlides.csv	Data for spots on all slides in APO knock-out expt as downloaded
13-6-data.gpr – 3-9-data.gpr	4 GenePix analysis results files
Contrasts13-6-9.gsh	Matrix holding contrast between GenePix treatments
Data13-6-9.gwb	Combined data set containing 4 GenePix slides
Slides13-6-9.gsh	Treatments on each of the 4 GenePix slides
Estimates13-6-9.gsh	Estimated effects for GenePix expt
ATH1-121501B.CDF	Chip information (layout and probes) for Affymetrix Arabidopis expt
Hyb1191.CEL – Hyb1400.CEL	Affymetrix files containing image analysis results for Arabidopis chips
Hyb-AllData.gwb	All data for Arabidopis expt merged into a single file
Hyb-PM_MM.gwb	Arabidopis data reorganised into PM/MM columns
Hyb-ANOVA.gwb	Results of ANOVA from Arabidopis expt
HybContrasts.gsh	Matrix holding contrasts between treatments for Arabidopis expt
Hyb-Expressions.gsh	Estimated expression values for Arabidopis expt
HybFiles.gsh	Description of 9 chips used in Arabidopis expt
Swirl1.csv – Swirl4.csv	Data on 4 slides for Zebra Fish Swirl expt as downloaded
Swirl_layout.csv	Layout of slides for Zebra Fish Swirl expt as downloaded
Swirl_layout.gwb	Layout of slides for Zebra Fish Swirl expt in Genstat format
SwirlSample.csv	Samples on slides for Zebra Fish Swirl expt as downloaded
SwirlSample.gsh	Samples on slides for Zebra Fish Swirl expt in Genstat format
Swirl.gsh	Combined data for Zebra Fish Swirl expt in Genstat format

You can also open these files using the File | Open Example Data Sets menu. If you select the Analysis of Microarray data filter, this will only show the files that are associated with this guide. If you do use the File | Open method to access the files, once you have navigated to the Microarrays folder, click the **Working directory Set as** button so that this will be the default directory accessed while working through the guide.



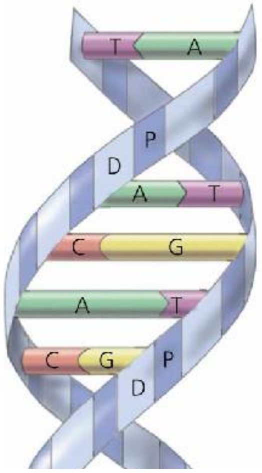

Example Data Sets

Look for file: APoAIGeneNames.tab

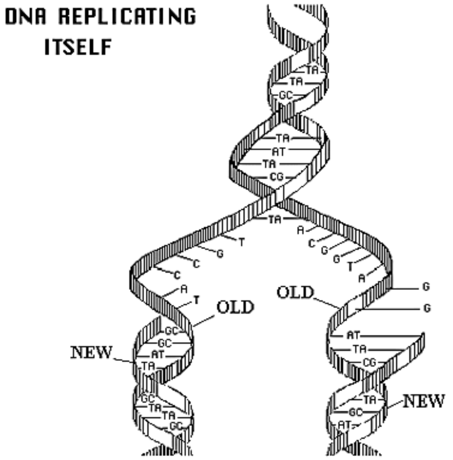
Filter by topic: Analysis of Microarray data

File	Description
APoAIGeneNam…	Gene names for APO knock-out mouse expt as downloaded from
APoAIGenes.gsh	Gene names for APO Knock-out mouse expt in Genstat format
ApoAIKnockOut.…	Data for spots on all slides in APO knock-out expt in unstacked f
ApoAIKnockOut…	Matrix holding contrast between knock-out & standard treatmen
ApoAIKnockOut…	Estimated effects for APO knock-out expt

Introduction



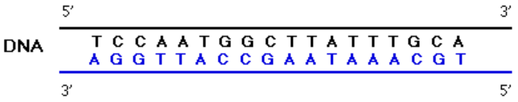
DNA is used to carry the instructions for cell processes. DNA is made up of four nucleotide bases: adenine, cytosine, guanine, and thymine (abbreviated as **A**, **C**, **G** and **T** respectively). These bases join into two complementary pairs, with A only binding to T and C with G. The bases are arranged in a double stranded helix (the backbone of the strands being made up of phosphate, the 5-carbon sugar deoxyribose) with complementary pairs of bases on each strand. Two single strands of DNA that have complementary bases at all matching positions (such as ACTGTGA and TGACACT) are known as complementary sequences. In a solution of the right temperature, these two single stands will bind together, to form a single double stranded section of DNA. If the temperature of the solution is raised the doubled stranded DNA will split back into two single stranded sections of DNA.



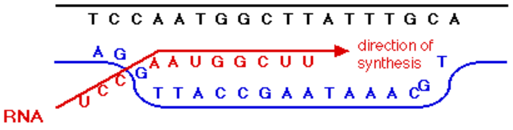
To reproduce, a cell must copy and transmit its genetic information (DNA) to its progeny. To do so, DNA replicates, following the process of semi-conservative replication. The two strands separate and each strand of the original molecule acts as a template for the synthesis of a new complementary DNA molecule.

DNA is a permanent/long term copy of the cell's information kept in the cell's nucleus. To express the information in the DNA, a single stranded copy of the bases is made. This single strand is known as RNA, and it uses the base uridine (**U**), in place of the base T. DNA serves as the template to make RNA. This process is known as **transcription** where information in the form of a sequence of bases is transferred from a double stranded DNA molecule to a single stranded RNA molecule, as shown to the right. Each group of three bases in RNA (a **codon**) code for a protein (see the diagram below right). The amino acids and the codons that code for each are given in Table 1.

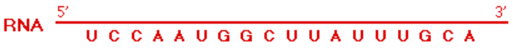
Transcription of RNA from DNA


5'
3'
DNA
T C C A A T G G C T T A T T T G C A
A G G T T A C C G A A T A A A C G T
3'
5'

- The bottom strand of the DNA molecule above is the template for RNA synthesis.
- RNA polymerase makes a copy of the DNA sequence but substitutes uridine (U) in place of thymine (T).


T C C A A T G G C T T A T T T G C A
A G
G
A A U G G C U U
direction of synthesis
T
U C C
T T A C C G A A T A A A C
G
RNA

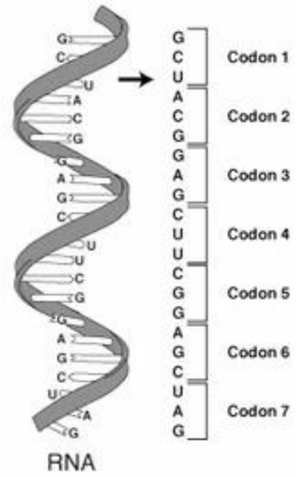
- The botttom strand of the DNA duplex is used as the template to synthesize RNA. However, the sequence of bases in the RNA is the same as in the top strand of the DNA, with U in place of T



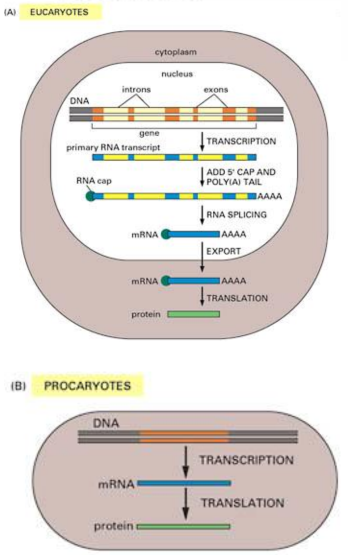
The RNA code is converted to a sequence of proteins in the ribosome in a process called **translation**. The section of the DNA that is transcribed as a unit is known as a **gene**, and starts with a start sequence (start codon, e.g., AUG or GUG) and finishes with an end sequence (stop codon, e.g., UAG etc).

Table 1: The 20 amino acids used in proteins and the codons that code for each amino acid.

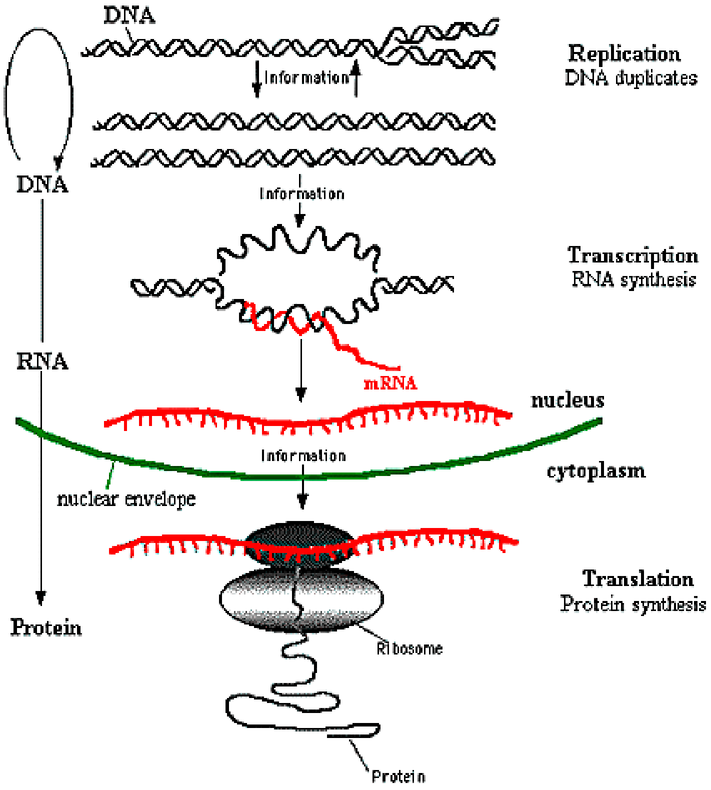
Amino acid	Code	Codon			
Ala	A	GCU, GCC, GCA, GCG	**Leu**	L	UUA, UUG, CUU, CUC, CUA, CUG
Arg	R	CGU, CGC, CGA, CGG, AGA, AGG	**Lys**	K	AAA, AAG
Asn	N	AAU, AAC	**Met**	M	AUG
Asp	D	GAU, GAC	**Phe**	F	UUU, UUC
Cys	C	UGU, UGC	**Pro**	P	CCU, CCC, CCA, CCG
Gln	Q	CAA, CAG	**Ser**	S	UCU, UCC, UCA, UCG, AGU, AGC
Glu	E	GAA, GAG	**Thr**	T	ACU, ACC, ACA, ACG
Gly	G	GGU, GGC, GGA, GGG	**Trp**	W	UGG
His	H	CAU, CAC	**Tyr**	Y	UAU, UAC
Ile	I	AUU, AUC, AUA	**Val**	V	GUU, GUC, GUA, GUG
Start		AUG, GUG	***Stop***		UAG, UGA, UAA



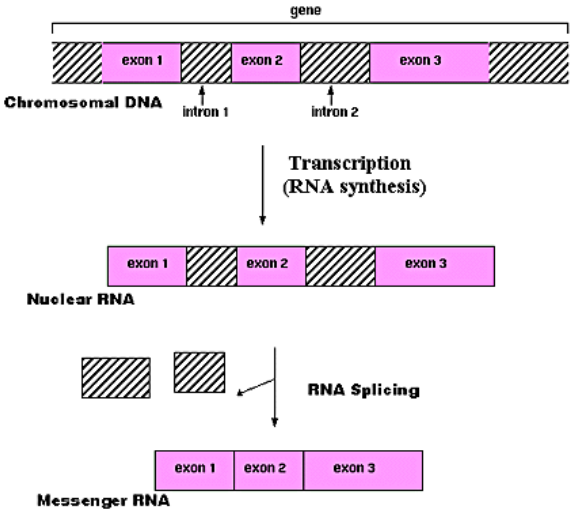
In eukaryotes (multi-cellar organisms which have a cell nucleus, as opposed to prokaryotes which are singular celled organisms with no nucleus, such as bacteria), not all the DNA is copied to the RNA, as some bases in so called introns are **spliced** out of the RNA after it is copied. The sections of DNA which code for the protein coding sections in this case are known as exons (see the diagram below right for details of introns/exons).



The central dogma of genetics is displayed in the diagram below. DNA is the basis of passing on the cellular information from one cell to another. DNA will replicate itself by each strand creating a new copy. In the nucleus, DNA is transcribed to RNA, which then moves out of the nucleus, where it is translated to proteins in the ribosomes. The expression of proteins controls the cell's mechanisms.



The Central Dogma of Molecular Biology

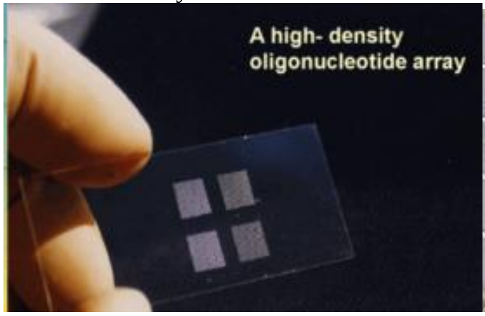


RNA synthesis and processing

Microarrays

A microarray is a glass slide or silicon chip that has had a library (a collection of DNA sequences) of single stranded DNA fragments laid down on its surface. The size of the libraries can be very large from 1000 to 150,000 on the highest density slides. The set of single stranded DNA fragments laid out on the surface are known as the probes, and these can be arrayed as a series of spots or squares in a grid depending on the procedure used to lay the DNA down. Robotic printing with pins or an inkjet printer gives rise to circular spots of DNA, and in situ creation of the DNA will create squares, as seen in Affymetrix chips. The DNA library can be created in various manners, as either complementary DNA (cDNA) that has been captured from cells as messenger RNA (mRNA) and sequenced to create expressed sequence tags (ESTs), or artificial sequences of DNA known as oligionucleotides (oligios). Oligios of a given length are often called n-mers, so for example an oligionucleotides of 25 bases as used on Affymetrix chips is a 25-mer.

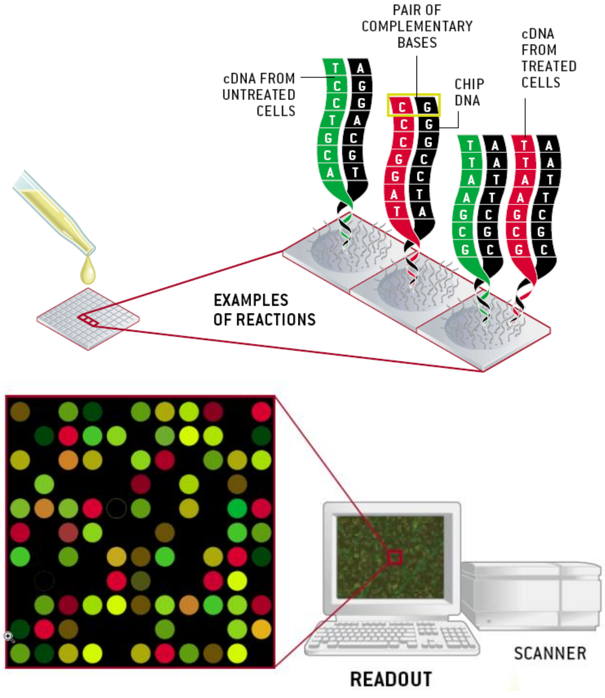
Glass Microarray slide



Robot printing microarrays

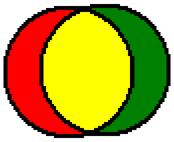
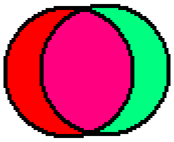
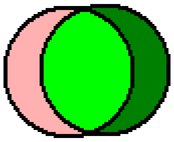
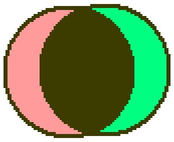
Affymetrix Silicon Chip

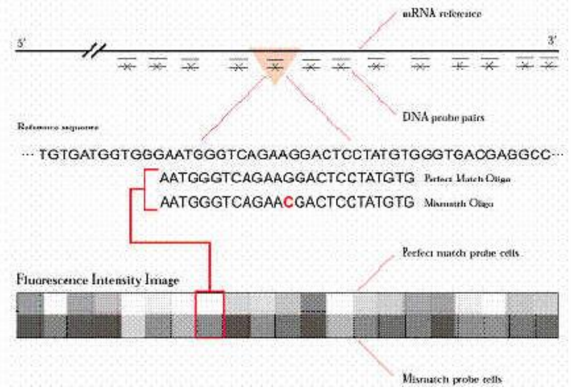
The microarray is then used to detect whether a sample of DNA or RNA contains sections of DNA that are complementary to the members of the library of DNA printed onto the slide. The DNA/RNA from the sample is known as the **target**. The target DNA is labelled with a fluorescent dye or an antibody that will bind to a dye. In two colour microarrays, two dyes, a red (**Cy5**) and a green (**Cy3**) dye are used, whereas in Affymetrix chips an antibody to biotin is used. The dyed samples are then added to the slide and left to allow the pairing of complementary sections of DNA in the target to bind to the probes on the slide (**hybridization**). After the DNA has bound, the slide is washed to remove unbound DNA, and the resulting levels of DNA bound to each spot are read off with a **laser scanner**. For two colour slides, two samples of RNA from different cell lines, cells under different conditions, different tissue types or individuals are added, one dyed red and the other green. In future three or more dyes may be able to be used to allow more samples per slide to be added. For the Affymetrix antibody-based staining, labelled biotin is added to the slide, and this will stick to the bound antibody on each spot. Various effects can cause errors in the level read to each spot, including non-specific hybridization where DNA that closely matches the probe also binds to the probe, and various artefacts and sources of noise that cause trends in the levels of the dyes across the microarray.



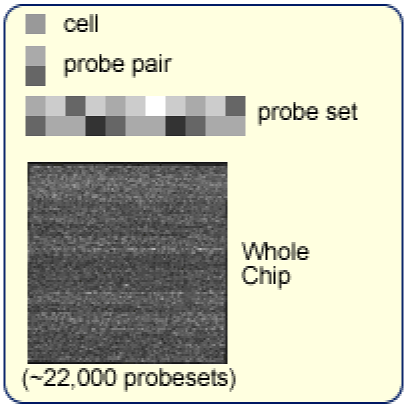
When the two levels of red and green are combined in a combined image, spots with high levels of red and green will display as yellow, spots with high red and low green as red, low red and high green as green and low in both red and green as black as shown below.

Table 2. The combinations of red and green dye levels displayed as in a microarray image.

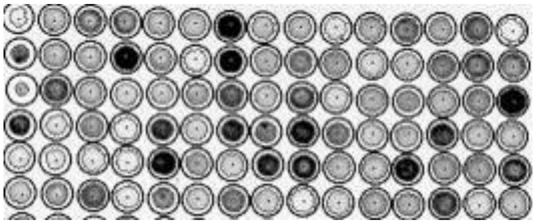
High Red & Green = Yellow/White	High Red & Low Green = Red	Low Red & High Green = Green	Low Red & Green = Black
			

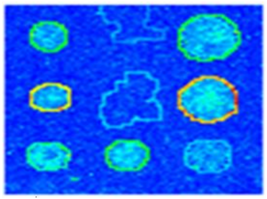

mRNA reference
5'
3'
DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC···
AATGGGTCAGAAGGACTCCTATGTG
Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTG
Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

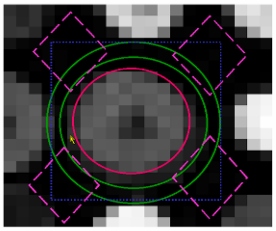
The layout of Affymetrix chips is more complex than other slides, as sets of probes are chosen for each gene of interest. Each gene is represented on the array by a series of different oligonucleotide probes. Each **probe pair** consists of a **perfect match** oligonucleotide (PM) and a **mismatch** oligonucleotide (MM). The perfect match probe has a sequence exactly complimentary to the particular gene and thus measures the expression of the gene. The mismatch probe differs from the perfect match probe by a single base substitution at the centre base position, disturbing the binding of the target gene transcript. This helps to determine the background and non-specific hybridization that contributes to the signal measured for the perfect match oligio. The GeneChip Operating Software MAS algorithm subtracts the hybridization intensities of the mismatch probes from those of the perfect match probes to determine the absolute or specific intensity value for each probe set. Probes are chosen based on current information from Genebank and other nucleotide repositories. The sequences on the expression arrays are believed to recognize unique regions of the 3' end of the gene. The diagrams to the right shows a schematic of a probe set.



The image files that are produced by the laser must be analysed by image analysis to give an intensity for each sample on every spot. This involves locating each spot of the slide, and then deciding which pixels should be read for that spot. The image to the right shows the estimated locations of the spots from one package. A wide range of image analysis packages used for this, including GenePix, Imagene, Spot. Each package uses its own algorithm, for example, Spot uses adaptive segmentation, where each spot can have its own shape and size which is estimated from the image (as shown in the coloured image to the right). The packages will also estimate the background level for each spot, and the image to the level shows the regions used for this for three packages. Each package tends to produce the results in its own format, and these have been allowed for in the Genstat menu that reads in the results.







Design of two-colour microarray experiments

In a two-colour microarray experiment, many slides are needed to allow assessment of several treatments. The design of the experiment is the choice of which pair of target samples (treatments) to put on each slide, and which colour to assign to each treatment in the pair. Technically, if the dye effects are estimated and removed, and since only the difference between the two samples is used from each slide (as we calculate log-ratios to estimate differential expression), the design is a row-column design with each row being a slide, and the two columns being the red and green dyes. Thus, obtaining an efficient row-column design for the number of target treatments will give an efficient microarray design. In terms of the efficiencies of particular treatment comparisons, the more often two treatments are on the same slide, the higher the precision of that comparison. However, even if two treatments do not appear on the same slide, they can still be compared indirectly through common treatments that appear with them both. For example, A and B can be compared if they both occur with C (A, C put on one slide and B, C on another). An indirect comparison has twice the variance of a direct comparison. The number of indirect comparisons typically grows much faster than the number of direct comparisons, and so can have a large impact on the efficiency of the design. If two treatments have no direct or indirect comparisons, then the design is said to be **disconnected**, as there are some comparisons that cannot be calculated from the design. A design may also be disconnected if there are no dye swaps in the trial, so that the dye and a treatment effect are confounded. In this case, you can drop the estimation of the dye effect, but this is in general poor practice.

Also, in terms of efficiency, if the slides are paired so that each pair has the same two treatments on each slide, but the two dyes are swapped between targets on the slides (a dye swap pair, e.g., **S**-**T** & **T**-**S**), then the treatment effects will be unconfounded with the dye effects (i.e. orthogonal to the dye effects). There is a less strict requirement for full efficiency of treatment estimation relative to dye effects, which is that each treatment only need to occur an equal number of times on each dye. There are several simple types of design that have been used in microarray experiments.

1. Reference designs

In a reference design, each treatment is compared with a standard treatment. In some reference designs, the standard is not even a treatment of interest, in which case the comparisons between the treatments are all **indirect**, being made only through the standard. A reference design is typically less efficient than other designs, especially if the standard is not of interest. A reference design may make sense if there are many treatments and there are only limited amounts of DNA available for each treatment.

Example reference design

This design compares B, C and D with the standard treatment A. Each comparison has a dye swap so that the design is balanced for dye effects.

Slide	Red	Green
1	A	B
2	B	A
3	A	C
4	C	A
5	A	D
6	D	A

2. Loop designs

In a loop design, the treatments flow from one slide on to another, with one treatment moving to the next slide, but changing dye, and a new one being introduced. When the treatments are exhausted, the treatment on the first slide is put on the final slide to provide a loop back to the first slide. For three treatments, this is equivalent to the balanced incomplete block design. One property of this design is that treatments must be balanced on the two dyes as each treatment occurs twice in each loop, one on each dye.

Example loop design

This experiment compares every treatment of A, B, C, D with every other treatment.

Slide	Red	Green
1	A	B
2	B	C
3	C	D
4	D	A

3. Balanced incomplete block designs

A balanced incomplete block design compares each treatment with every other treatment an equal number of times. Due to the high level of linking between treatments, the number of indirect comparisons grows very quickly in a balanced incomplete block design, normally leading to high efficiencies. In addition, every treatment comparison has the same level of precision.

Example balanced incomplete block design

This experiment compares every treatment of A, B, C, D with every other treatment. Note that as each treatment occurs 3 times, they cannot be completely balanced for dye, but are even as possible occurring once on one dye and twice on the other.

Slide	Red	Green
1	A	B
2	C	A
3	A	D
4	B	C
5	D	B
6	C	D

Structured treatments

If there is a higher order structure to the treatment targets, then estimates of particular comparisons between the treatments (contrasts) can be made. The contrasts give the coefficients of the treatment means when they are summed to give the summary statistic. Typically, as we are interested in differences between treatments, the sum of the contrast coefficients is zero. For example, if we were interested in the difference between the mean of two treatments A & B and two other treatments C & D, we would calculate this difference (A + B)/2 – (C + D)/2 so that the contrast coefficients would be (½, ½, -½, -½). Often, we work with a multiple of the contrast to avoid fractions, so we could use the equivalent contrast for this of (1, 1, -1, -1). The common treatment structures that are used are treatments associated with a quantitative measure (time, concentration of chemical etc) and **factorial** combinations of two or more terms (e.g., cell type by cell age, animal line by experimental intervention). If we want to explore changes with the quantitative treatment and the differential expression, often polynomial contrasts are used to measure linear trend, curvature (the quadratic component) etc.

Examples of contrast matrices

The following illustrate three types of contrasts and the matrices set up to estimate these.

Comparing the mean of two treatments with another treatment

Treatment	A	B	C	D
A vs. B, C	-2	1	1	0

2 x 2 Factorial treatments structure - main effects and interaction

Treatment	A_1B_1	A_1B_2	A_2B_1	A_2B_2
A main effect	-1	-1	1	1
B main effect	-1	1	-1	1
A×B interaction	-1	1	1	-1

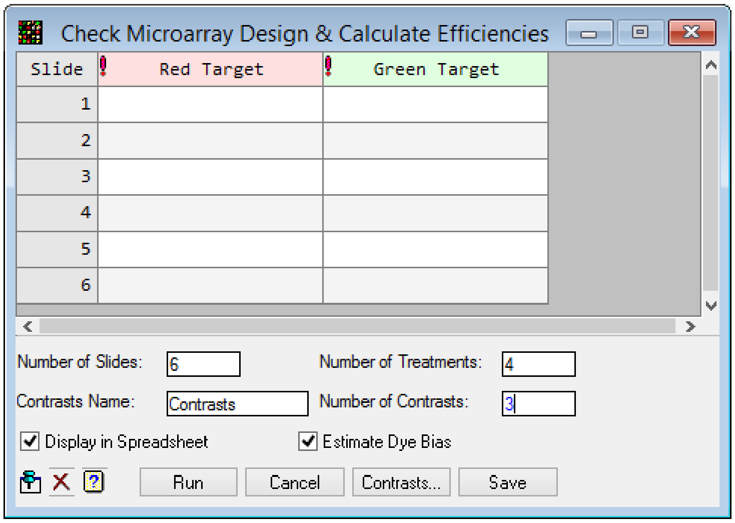
Polynomial contrast for 4 treatments with uniform spacing

Treatment	T_1	T_2	T_3	T_4
Linear	-3	-1	1	3
Quadratic	1	-1	-1	1
Cubic	-1	3	-3	1

The Genstat procedure **ORTHPOLYNOMIAL** can be used to generate the values needed to define a set of orthogonal polynomials with any spacing of points.

Example using the Microarray Design menu

The menu Stats | Microarray | Design | Check Two Channel Design opens a dialog that allows you to examine the precision of the treatment comparison and optionally creates contrasts for a given design. The number of treatments, slides and contrasts are entered into the dialog, and then the targets for each slide and dye are entered. Any labels can be used for the treatments, but the dialog checks that the number of distinct labels equals that specified in the **Number of Treatments** edit box. You need to be careful to keep the case the same between cells, as B and b are taken as separate treatments. To save typing, cell labels can be put on the clipboard with Ctrl+C and pasted back to a new cell with Ctrl+V. Multiple cells can be selected by holding down the Shift key while using the arrow keys or clicking with the mouse in a cell and dragging the selection over the required cells.


Check Microarray Design & Calculate Efficiencies

Slide	Red Target	Green Target
1		
2		
3		
4		
5		
6		

Number of Slides: 6

Number of Treatments: 4

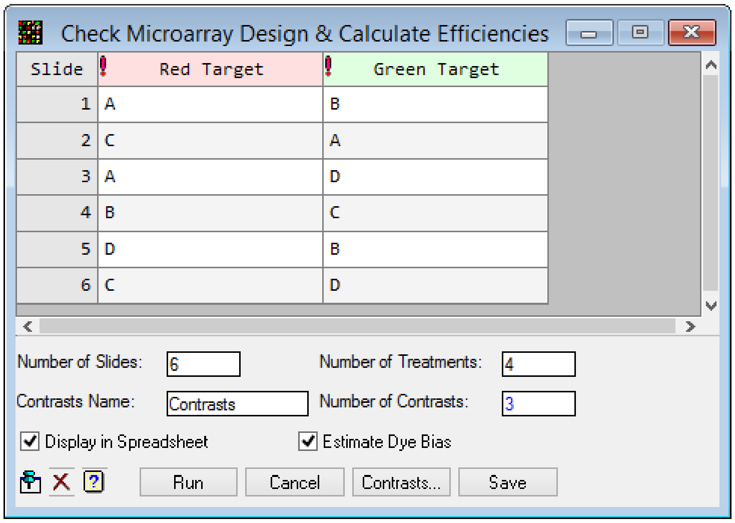
Contrasts Name: Contrasts

Number of Contrasts: 3

☑ Display in Spreadsheet

☑ Estimate Dye Bias

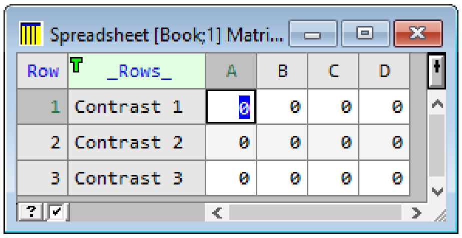
For example, if we enter three replicates for four treatments (A, B, C, D) in an incomplete block design, we will complete the dialog as shown to the left. Notice that with three replicates of each treatment, we cannot get dye balance, but each treatment is used twice with one dye and once on the other, making the design as balanced as possible. If we have a series of contrasts between the four treatments, we fill in the name of a matrix in the **Contrasts Name** edit box, and the number of rows in the matrix in the **Number of Contrasts** edit box. The matrix has a row for each contrast, and a column for each treatment. If the matrix does not exist, we can create this by clicking the **Contrasts** button, which creates a matrix with the columns labelled by the treatment labels used in the design spreadsheet and with the number of rows specified. If the name of the contrast matrix or the number of rows has not been entered, you will be prompted for these.


Check Microarray Design & Calculate Efficiencies

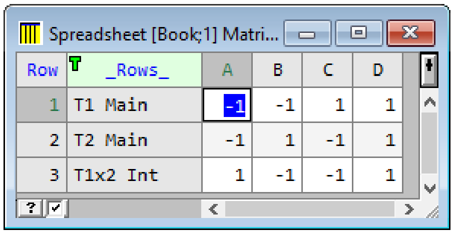
Slide	Red Target	Green Target
1	A	B
2	C	A
3	A	D
4	B	C
5	D	B
6	C	D

Number of Slides: 6
Number of Treatments: 4
Contrasts Name: Contrasts
Number of Contrasts: 3
☑ Display in Spreadsheet
☑ Estimate Dye Bias

If the four treatments A-D were a 2 x 2 factorial combination of two 2 level treatments, we could create a matrix to estimate the main effects of both treatments and the two-way interactions as follows. Initially, the matrix would have default values as shown to right, and then you would complete it to obtain the values as shown to the far right.

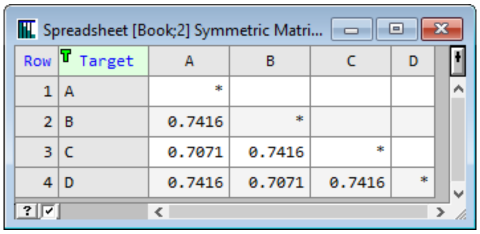

Spreadsheet [Book;1] Matri...

Row	_Rows_	A	B	C	D
1	Contrast 1	0	0	0	0
2	Contrast 2	0	0	0	0
3	Contrast 3	0	0	0	0

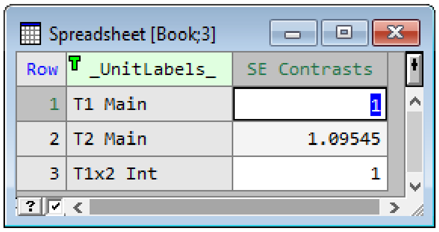

Spreadsheet [Book;1] Matri...

Row	_Rows_	A	B	C	D
1	T1 Main	-1	-1	1	1
2	T2 Main	-1	1	-1	1
3	T1x2 Int	1	-1	-1	1

Clicking the **Run** button will generate some output in the Output window and the following two spreadsheets, as the Display in spreadsheet option was set.


Spreadsheet [Book;2] Symmetric Matri...

Row	Target	A	B	C	D
1	A	*			
2	B	0.7416	*		
3	C	0.7071	0.7416	*	
4	D	0.7416	0.7071	0.7416	*


Spreadsheet [Book;3]

Row	_UnitLabels_	SE Contrasts
1	T1 Main	1
2	T2 Main	1.09545
3	T1x2 Int	1

The first spreadsheet gives the precision of each pairwise comparison between the four treatments. The leading diagonal is missing, as you do not want to compare a treatment with itself. The comparisons A-C and B-D are slightly more accurate that the others, because A and C have the same pattern across Red and Green (2-1) and B and D having the other pattern (1-2). However, the variation in SEDs is quite small (< 5%). Note these SEDs assume a variance of 1 between slides, and so in the actual experiment, the SEDs will depend on the achieved error variance between slides. The second spreadsheet gives the precision of the contrast that were set up in the matrix above. The T2 main effect is not as precisely estimated as the T1 main effect and the interaction. The output in the Output window is displayed on the next page.

Microarray design efficiencies

Microarray design

```
   Slide     Red_Trt    Green_Trt
       1           A            B
       2           C            A
       3           A            D
       4           B            C
       5           D            B
       6           C            D
```

The design is printed out again for reference.

Variance covariance matrix

```
   A     0.2000
   B    -0.0750     0.2000
   C    -0.0500    -0.0750     0.2000
   D    -0.0750    -0.0500    -0.0750     0.2000
 Dye    -0.0500     0.0500    -0.0500     0.0500     0.2000
              A          B          C          D        Dye
```

The variance-covariance matrix shows confounding between treatment terms. Here for example the Dye – treatment terms of 0.05 show that the treatments are not balanced on the two dyes.

Standard errors of differences

```
  A           *
  B      0.7416          *
  C      0.7071     0.7416          *
  D      0.7416     0.7071     0.7416          *
              A          B          C          D
```

This repeats the information in the SED spreadsheet.

Summary statistics of SEDs

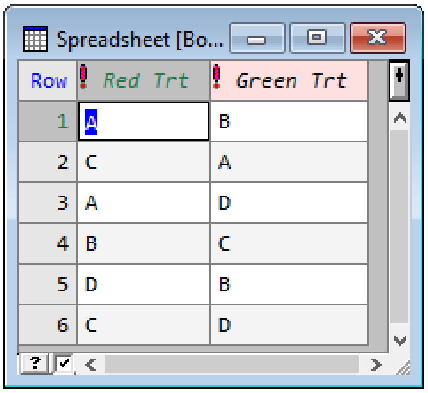
```
     Minimum           Mean    Maximum
      0.7071         0.7301     0.7416
```

This gives a summary over all the SEDs.

Standard errors of contrasts

```
     T1 Main     1.0000
     T2 Main     1.0954
     T1x2 Int    1.0000
```

This repeats the information displayed in the SED of contrasts matrix.



Row	Red Trt	Green Trt
1	A	B
2	C	A
3	A	D
4	B	C
5	D	B
6	C	D

Once you have a design that you are happy with you can save it to a spreadsheet, by clicking the **Save** button. This will create the spreadsheet to the right. You can save this to disk with the File | Save menu. This can be copied back to the design dialog later by selecting all the cells (Ctrl+A) and copying them with Edit | Copy (Ctrl+C). Back in the design dialog, set the correct number of treatments and slides, and then paste the labels in with the Ctrl+V key combination (note that you cannot use the menu bar for this, but must use the keyboard).

Blocking and randomization in microarray designs

When running the slides in a microarray experiment, it is possible to group the slides so that they fall into blocks, ensuring that if there is variation throughout the experiment due to differences between operators, days, lab kits, printing variation on the slides, then these effects will be balanced out over the treatments. As always, it is best to include some randomization in the order of processing the slides so that treatment effects are not confounded with trends over the trial.

The design to the right has 3 treatments, each containing 2 individual animals. The microarray experiment is designed in 3 replicates, and within each replicate, the slides have been grouped into blocks of 3 slides. Each block contains dye-balanced comparisons between A-B, A-C, B-C with each animal used once. The replicates contain dye-balanced comparisons of every animal compared with every other animal in the other treatments, (e.g., A1 is compared with B1, B2, C1 and C2). Over the 3 replicates, the individual animal comparisons are as dye balanced as possible (A1-B1 occurs twice and B1-A1 occurs once). Note in the figure, the treatment assigned to the red dye is in the left-hand column and that assigned to the green dye in the right-hand column.

Rep 1		**Rep 2**		**Rep 3**	
A1	**C2**	**C2**	**A1**	**A1**	**B2**
B2	**A2**	**A2**	**B2**	**B1**	**C1**
C1	**B1**	**B1**	**C1**	**C2**	**A2**
A2	**C2**	**C2**	**A2**	**A1**	**C1**
B1	**A1**	**A1**	**B1**	**C2**	**B1**
C1	**B2**	**B2**	**C1**	**B2**	**A2**
A1	**C1**	**C1**	**A1**	**B1**	**A1**
C2	**B2**	**B2**	**C2**	**A2**	**C2**
B1	**A2**	**A2**	**B1**	**C1**	**B2**
B2	**A1**	**A1**	**B2**	**A2**	**B1**
C2	**B1**	**B1**	**C2**	**B2**	**C2**
A2	**C1**	**C1**	**A2**	**C1**	**A1**

Microarray design exercises

The following exercises can be attempted with Microarray Design menu. The solutions are given on the following page.

1. With 12 slides produce at design with 4 treatments that maximizes the overall precision of the treatment comparisons.

2. Compare the design you have found with a standard reference design (all treatments vs. a common control where a, the control is treatment 1, and b, the control is not one of the treatments), and with the loop design.

3. If the 4 treatments represent a 2 x 2 factorial design, find the design which maximizes the precision of estimating the interaction (assuming treatments 1-4 are A_1B_1, A_1B_2, A_2B_1, and A_2B_2, contrast levels for A & B main effects and interaction are (-1, -1, 1, 1), (-1, 1, -1, 1) and (1, -1, -1, 1).

4. If the 4 treatments represent 4 times in a time course experiment, find a design optimizes the estimation of the linear effect (contrast levels = (-3, -1, 1, 3)) while still allowing estimation of the changes between adjacent times. Which arrangement of a loop design makes the linear contrast the most accurate?

5. If you can add 2 more slides to the designs in 1 and 3, which allocation to the 2 slides would you use?

6. If you have 2 treatments each with 4 animals, design a trial with 8 slides which optimizes the calculation of the animal-animal variation, while optimally measuring the between treatment difference. Can you make this design balanced with respect to dye swaps for each animal? Also, design a trial with 12 slides, and compare the gain in precision between the 2 designs.

Solutions to design exercises

1. The optimal design to give the largest average standard error of a difference is a balanced incomplete block design, which compares every treatment with every other one twice. The second replicate should be a dye swap of the first to obtain dye balance. Every comparison between treatments has the same standard error of 0.5

Red	Green
A	B
B	C
C	D
D	A
A	D
D	C
C	B
B	A
A	C
C	A
B	D
D	B

SEDs of Treatment Comparisons				
Target	**A**	**B**	**C**	**D**
A	*			
B	0.5	*		
C	0.5	0.5	*	
D	0.5	0.5	0.5	*

2. The standard reference design is given below. The comparisons with A have the same precision as the incomplete block design, but the indirect comparisons between B, C and D have a lower precision (0.7071).

Target	A	B	C	D
A	*			
B	0.5	*		
C	0.5	0.7071	*	
D	0.5	0.7071	0.7071	*

Red	Green
A	B
A	C
A	D
B	A
C	A
D	A
A	B
A	C
A	D
B	A
C	A
D	A

The loop design, using the order A, B, C and D is given to the right. This has 3 loops, one that is a dye swap of the other two. This has adjacent treatments compared with the same accuracy as the balanced incomplete block design, and a slightly lower precision for the other comparisons (0.5774)

Target	A	B	C	D
A	*			
B	0.5000	*		
C	0.5774	0.5000	*	
D	0.5000	0.5774	0.5000	*

Red	Green
A	B
B	C
C	D
D	A
B	A
C	B
D	C
A	D
A	B
B	C
C	D
D	A

3. The previous loop design optimizes the interaction precision if the substitution A = A_1B_2, B = A_1B_1, C = A_2B_1, and D = A_2B_2 is made. The precision of the main effect and interaction contrast as given in the following table.

Effect	SE Contrasts
A	0.8165
B	0.8165
AB	0.5774

4. The allocation of A, B, C, and D that optimises the linear contrast precision is the one that maximizes the changes between adjacent times. This gives A, B, C, D = Times 2, 3, 1, 4 with differences of 1,2,3,2 as you run around the loop (returning from 4 to 2). The standard error of the linear contrast with this ordering is 1.414 whereas using the natural ordering of 1,2,3,4 with differences of 1,1,1,3 gives a standard error of 1.732, and the worst order of 2,1,3,4 with differences of 1,2,1,2 gives a standard error of 1.826.

5. In design 1, any two independent allocations (say A-B and C-D) will give an optimal design. The two comparisons represented on the extra slides will have a higher precision that the rest. In design 3, the extra slides $A_1B_1 – A_2B_1$, and A_1B_2 - A_2B_2 give the best standard error for the interaction of 0.535.

6. The only possible connected design is a loop through all the animals, alternating between the two treatments. The individual animals are not compared very precisely, with the SED matrix shown below. The contrast matrix of (-0.25, -0.25, -0.25, -0.25, 0.25, 0.25, 0.25, 0.25) for the mean of B – A has a standard error or 0.354. Adding 4 extra slides allows more accurate comparisons of individuals, and the 4 slides shown below reduce the between-animal comparisons average SED from 1.210 to 0.838 (and the maximum from 1.414 to 0.913). The treatment standard error is reduced to 0.289, an 18% reduction.

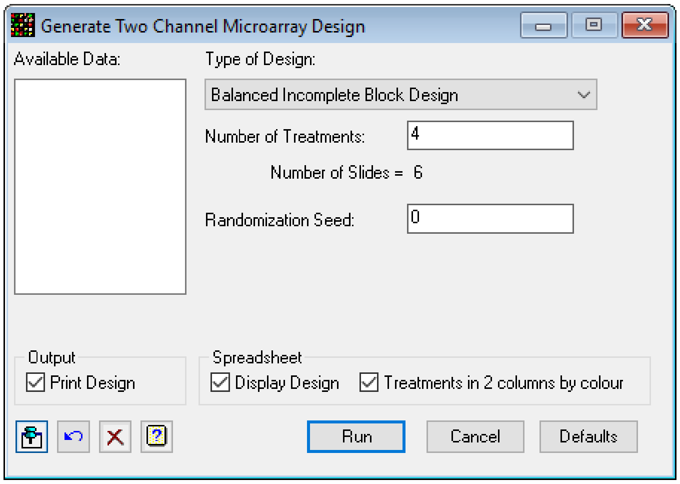
Red	Green
A1	B1
B1	A2
A2	B2
B2	A3
A3	B3
B3	A4
A4	B4
B4	A1

Target	A1	A2	A3	A4	B1	B2	B3	B4
A1	*							
A2	1.225	*						
A3	1.414	1.225	*					
A4	1.225	1.414	1.225	*				
B1	0.935	0.935	1.369	1.369	*			
B2	1.369	0.935	0.935	1.369	1.225	*		
B3	1.369	1.369	0.935	0.935	1.414	1.225	*	
B4	0.935	1.369	1.369	0.935	1.225	1.414	1.225	*

Added Slides	
A1	B2
B3	A2
A3	B4
B1	A4

Automatic generation of designs

The Stats | Microarrays | Design | Generate Two Channel design menu allows you to automatically generate reference, loop, and balanced incomplete block designs. You just need to specify the type of design, the number of treatments, and either the loop increment (1 always works), or the reference level. With a loop design, you can specify more than one loop increment to get two complementary loops. Using the **Treatments in 2 columns by colour** option puts the design in a format appropriate for pasting into the Stats | Microarrays | Design | Generate Two Channel design menu. The spreadsheet to the right shows the resulting balanced incomplete block design with 4 treatments.



Generate Two Channel Microarray Design

Available Data:

Type of Design: Balanced Incomplete Block Design

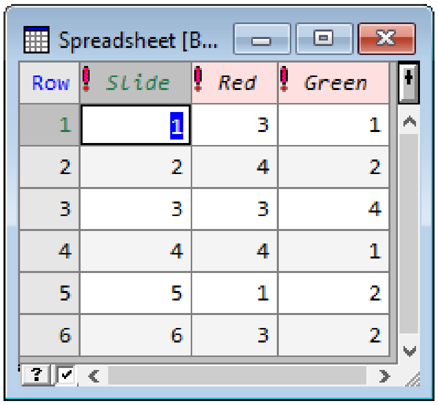
Number of Treatments: 4

Number of Slides = 6

Randomization Seed: 0

Output
☑ Print Design

Spreadsheet
☑ Display Design ☑ Treatments in 2 columns by colour

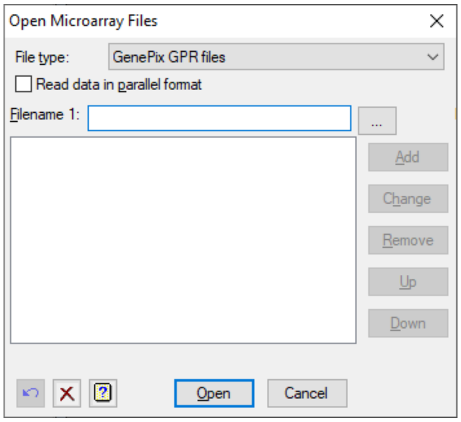


Spreadsheet [B...

Row	Slide	Red	Green
1	1	3	1
2	2	4	2
3	3	3	4
4	4	4	1
5	5	1	2
6	6	3	2

For single channel designs, the Stats | Microarrays | Design | Generate Single Channel design menu just opens the Generate Standard Design menu, as there are no complications with these designs, as each slide is an independent unit, as in any standard experimental design.

Reading microarray data


Open Microarray Files

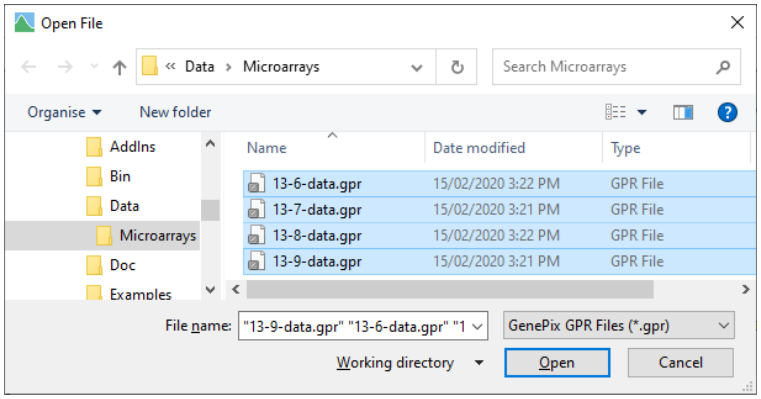
File type: GenePix GPR files

□ Read data in parallel format

Filename 1:

Microarray data can be read in via the usual File | Open menu in Genstat if it is in one of the standard file formats supported by Genstat (Excel, CSV, text etc). However, in microarray experiments you often have a file for each slide in the experiment and need these amalgamated into a single data set.

The Microarrays Data menu has a dialog that is designed for this, which also supports the common formats produced by image analysis software (GenePix, GenePix, Imagene, Spot, TIGR MEV, ScanAlyze, QuantArray, Affymetrix and generic CSV files). The Data menu contains a list of the supported file types, so chose the file type you have. These all open the same dialog, but just pre-select type of file on this menu (note if you have not set the working directory as explained on page 2, you will need to navigate to the microarray files in the Genstat Data\Microarrays folder). You can change between file types once the dialog is open. The selected files should all be of the same format, but some may have extra columns, in which case missing values will be inserted in the files with these columns missing. The file names may be typed in and added to the file list using the Add button, but more commonly the browse button (...) is used.


Open File

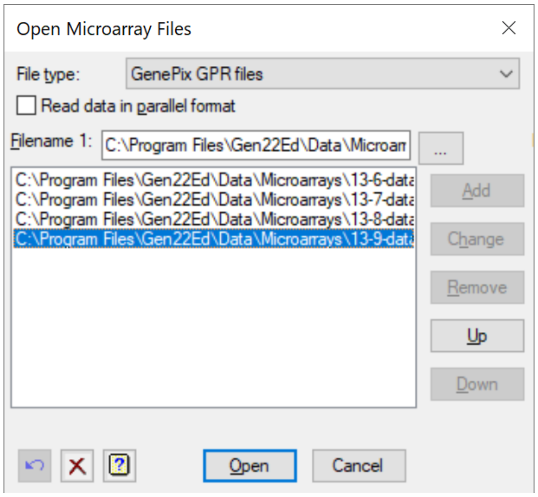
« Data › Microarrays

Name	Date modified	Type
13-6-data.gpr	15/02/2020 3:22 PM	GPR File
13-7-data.gpr	15/02/2020 3:21 PM	GPR File
13-8-data.gpr	15/02/2020 3:22 PM	GPR File
13-9-data.gpr	15/02/2020 3:21 PM	GPR File

File name: "13-9-data.gpr" "13-6-data.gpr" "1

GenePix GPR Files (*.gpr)

For example, to read in the 4 GenePix GPR files in the Ma2Examples directory, use the menu Stats | Microarray | Data | GenePix GPR files and click the browse button as shown. Now select the 4 GPR files required in the Select Microarray files dialog (as shown below, right), and click **Open**. To select multiple files, hold down the control key (Ctrl) or to select a range hold down the Shift key when clicking with the mouse. Note Windows (before version 7) does not enter the filenames selected in a natural order in the Filename box, so the resulting file names in the Open Microarray Files window are not in numerical order (as below). To arrange the files in the correct order, select the file to move (in this case “13-9-data.gpr”) and use the **Up** or **Down** button to move it within the list (**Down** 3 times in this case to put it at the end of the list).


Open Microarray Files

File type: GenePix GPR files

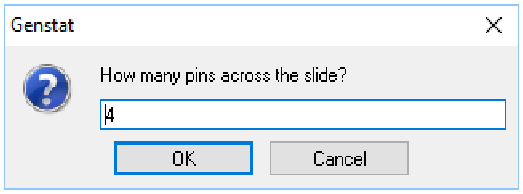
□ Read data in parallel format

Filename 1: C:\Program Files\Gen22Ed\Data\Microarr

C:\Program Files\Gen22Ed\Data\Microarrays\13-6-data
C:\Program Files\Gen22Ed\Data\Microarrays\13-7-data
C:\Program Files\Gen22Ed\Data\Microarrays\13-8-data
C:\Program Files\Gen22Ed\Data\Microarrays\13-9-data

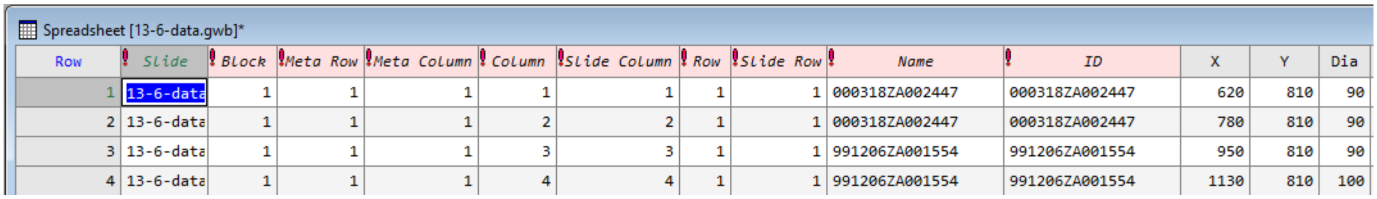
The option to read data in parallel format creates a new column for each slide rather than appending slides and uses a pointer to these columns for each slide which be used in menus that allow a pointer data format.

On clicking **Open**, the files are appended into a combined spreadsheet with an extra factor created to index the file the data came from. You may be prompted for some extra information, such as which columns are factors, dates and in the case of a GPR file, the arrangement of pins on the slide, as this information is not within the data. The final spreadsheet is shown on the next page.


Genstat

How many pins across the slide?

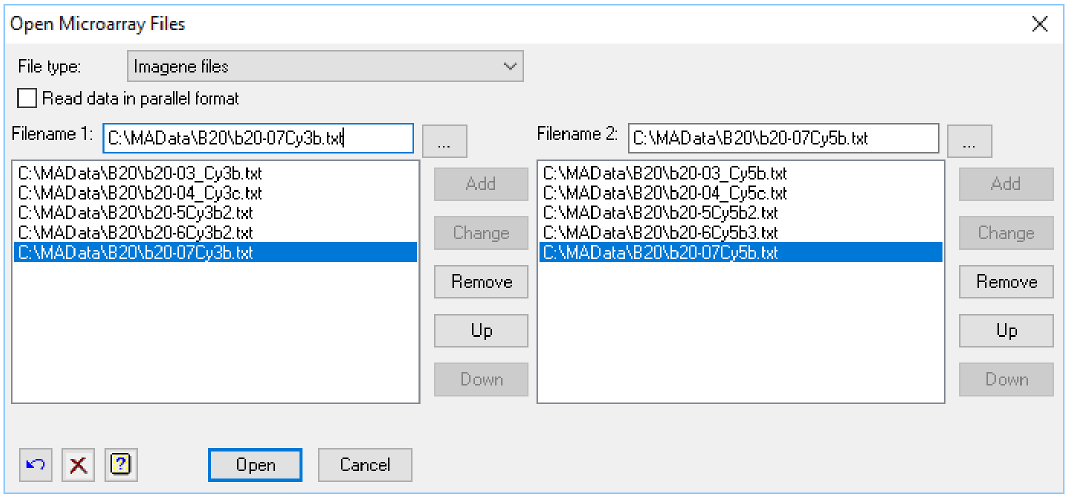
4



Spreadsheet [13-6-data.gwb]*

Row	Slide	Block	Meta Row	Meta Column	Column	Slide Column	Row	Slide Row	Name	ID	X	Y	Dia
1	13-6-data	1	1	1	1	1	1	1	000318ZA002447	000318ZA002447	620	810	90
2	13-6-data	1	1	1	2	2	1	1	000318ZA002447	000318ZA002447	780	810	90
3	13-6-data	1	1	1	3	3	1	1	991206ZA001554	991206ZA001554	950	810	90
4	13-6-data	1	1	1	4	4	1	1	991206ZA001554	991206ZA001554	1130	810	100

With Imagene files, the red and green channels are stored in two files. To read in this data you will need to give two lists of files, one for the red channel and one for the green channel. The order of files should be such that files corresponding to the same slide are in the same position in the list. The dialog below shows Imagene files being selected for input.



Open Microarray Files

File type: Imagene files

□ Read data in parallel format

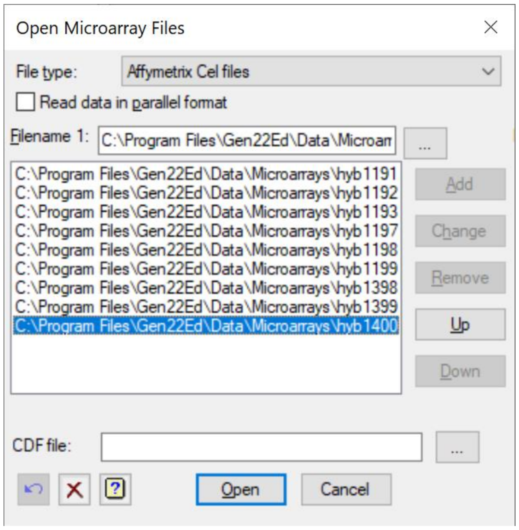
Filename 1: C:\MADatalB20\b20-07Cy3b.txt

C:\MADatalB20\b20-03_Cy3b.txt
C:\MADatalB20\b20-04_Cy3c.txt
C:\MADatalB20\b20-5Cy3b2.txt
C:\MADatalB20\b20-6Cy3b2.txt
C:\MADatalB20\b20-07Cy3b.txt

Filename 2: C:\MADatalB20\b20-07Cy5b.txt

C:\MADatalB20\b20-03_Cy5b.txt
C:\MADatalB20\b20-04_Cy5c.txt
C:\MADatalB20\b20-5Cy5b2.txt
C:\MADatalB20\b20-6Cy5b3.txt
C:\MADatalB20\b20-07Cy5b.txt

With Affymetrix CEL files, the option of merging the chip layout data that is stored in CDF file format is available. When opening cell CEL files, you also get the option of what columns are read in (to save space as CEL files are very large), and then the option to batch process the results rather than opening them into a spreadsheet, which is much faster and more memory efficient. The Affymetrix CEL Read Options dialog opens to specify these options. The Affymetrix analysis methods are described in the next section.



Open Microarray Files

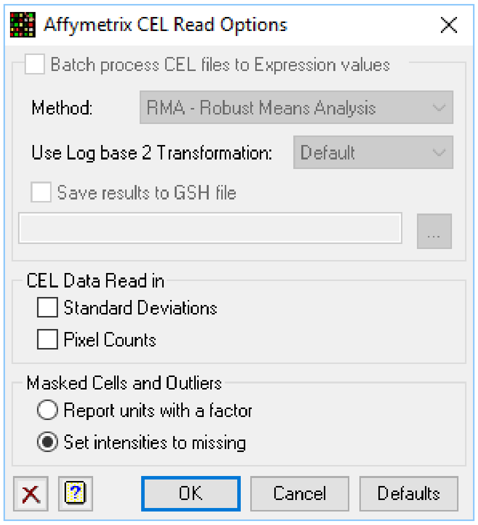
File type: Affymetrix Cel files

□ Read data in parallel format

Filename 1: C:\Program Files\Gen22Ed\Data\Microarr

C:\Program Files\Gen22Ed\Data\Microarrays\hyb1191
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1192
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1193
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1197
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1198
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1199
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1398
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1399
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1400

CDF file:



Affymetrix CEL Read Options

□ Batch process CEL files to Expression values

Method: RMA - Robust Means Analysis

Use Log base 2 Transformation: Default

□ Save results to GSH file

CEL Data Read in

□ Standard Deviations

□ Pixel Counts

Masked Cells and Outliers

○ Report units with a factor

● Set intensities to missing

Once the microarray data are in a spreadsheet, they can be plotted and manipulated as with any other spreadsheet in Genstat. The file can be saved permanently in Genstat Spreadsheet or Workbook format using the File | Save As menu. Note that the GWB format is more compact than the GSH format.

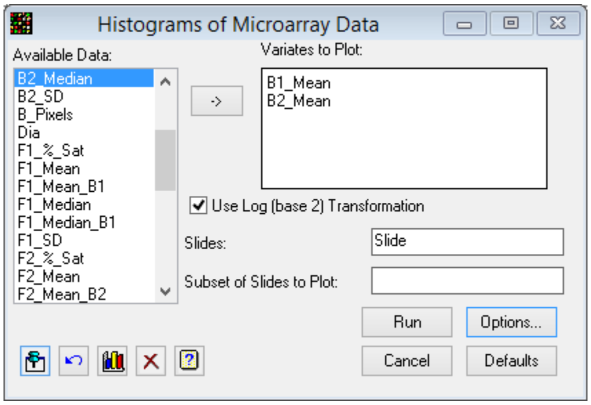
The Spread menu can be used to sort, filter, and edit the data, and the Graphics menu can be used to plot the data. There are some specialized Graphics menus for microarray in the Microarray | Explore menu, which is explained next.

Exploration of microarray data

The Explore submenu of the Microarray menu allows various customized plots of microarray data. The 4 explore menus are Histograms, Density, 2D Plots and Spatial Plot. These are standard Genstat graphics types, which may be found under the Graphics menus, but the microarray menus are customized to allow separate graphs for each slide, with subsets of slides to be displayed being able to set, as well as doing the data manipulation into the format required for the graphs.

Explore histogram

The Histogram menu allows you to look at the distribution of the measurements from several slides in the experiment to look for odd slides that have a differential response to the other slides. For example, if we open the file “Data13-6-9.gwb” that was saved from the 4 GenePix files opened in the previous example, then we can plot the histograms of various measurements for each slide. To look at the quality and level of the slide backgrounds, we could plot the variates B1_Mean and B2_Mean, which are the mean levels of red and green respectively around each spot. The menu allows either log$_2$ transformed or non-transformed values to be plotted, and log$_2$ values have been selected for the following plots. Completing the Histogram menu as shown to the right and using the options button to request multiple histograms per page as to the right, gives the plots below.


Histograms of Microarray Data

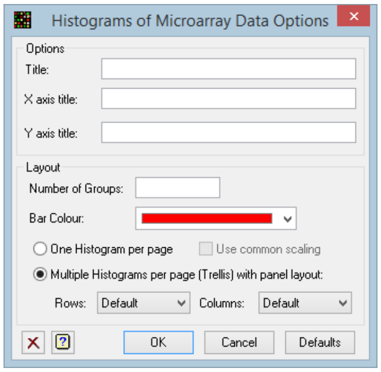
Available Data:
B2_Median
B2_SD
B_Pixels
Dia
F1_%_Sat
F1_Mean
F1_Mean_B1
F1_Median
F1_Median_B1
F1_SD
F2_%_Sat
F2_Mean
F2_Mean_B2

Variates to Plot:
B1_Mean
B2_Mean

☑ Use Log (base 2) Transformation

Slides: Slide

Subset of Slides to Plot:


Histograms of Microarray Data Options

Options

Title:

X axis title:

Y axis title:

Layout

Number of Groups:

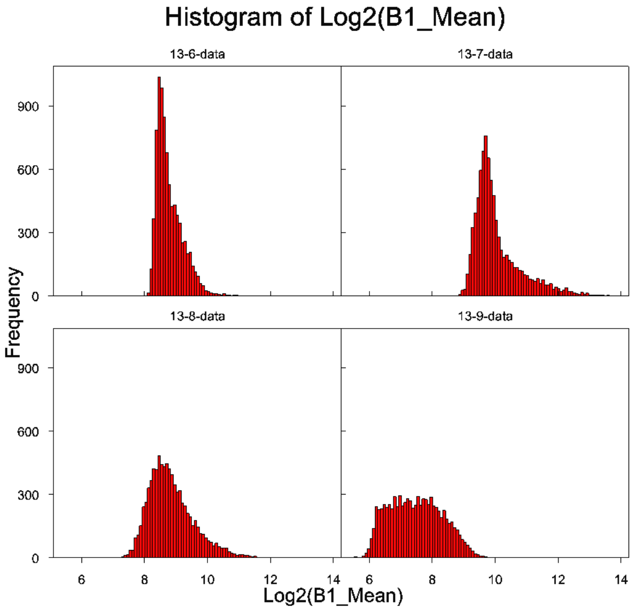
Bar Colour: 🟥

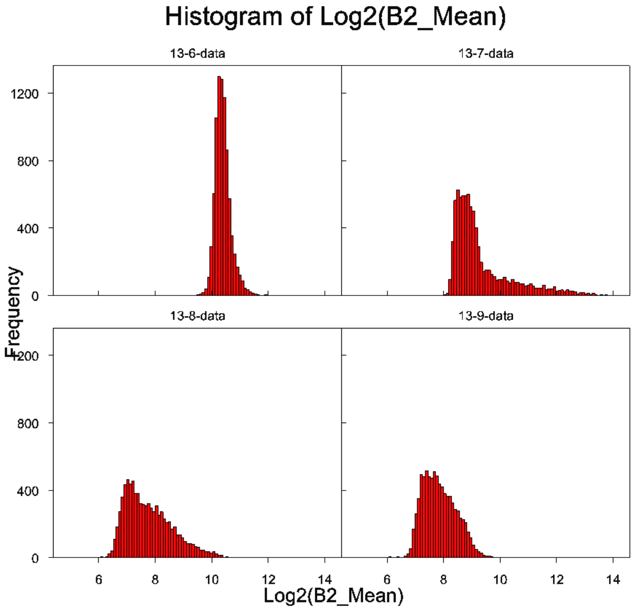
○ One Histogram per page □ Use common scaling

● Multiple Histograms per page (Trellis) with panel layout:

Rows: Default Columns: Default

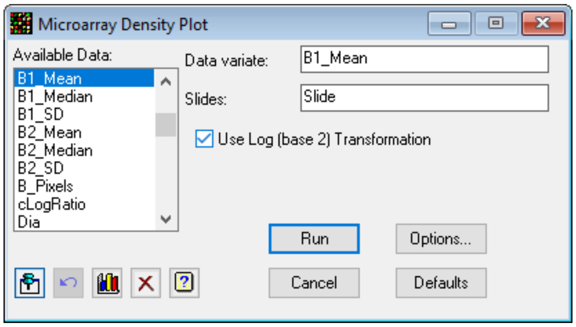
From these, it can be seen that the background levels on slide 13-6 cover a narrower range than the other slides, and that there is a higher background level on slide 13-7, with a long tail of very high backgrounds around some spots. Slide 13-9 has lower levels of background than the other slides, but a broader distribution. The red backgrounds tend to be lower than the green ones.





Explore density

The Density menu allows you to look at the distribution of the measurements as either a kernel smoothed probability density or as a cumulative density function. The first is like the Histogram display, but the slides can be plotted on the same graph for easier comparisons between slides. For example, using that data from the Histogram example above, and plotting the variate B1_Mean (the mean background levels of red) with the Density menu, completed as shown to the right, and with options set as right, gives the following plots.



Microarray Density Plot

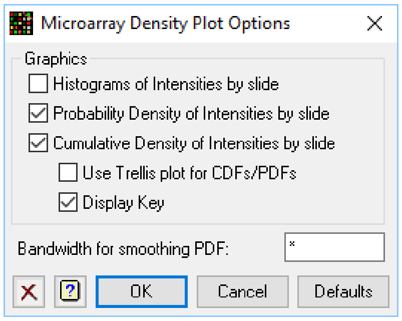
Available Data:

B1_Mean
B1_Median
B1_SD
B2_Mean
B2_Median
B2_SD
B_Pixels
cLogRatio
Dia

Data variate: B1_Mean

Slides: Slide

☑ Use Log (base 2) Transformation

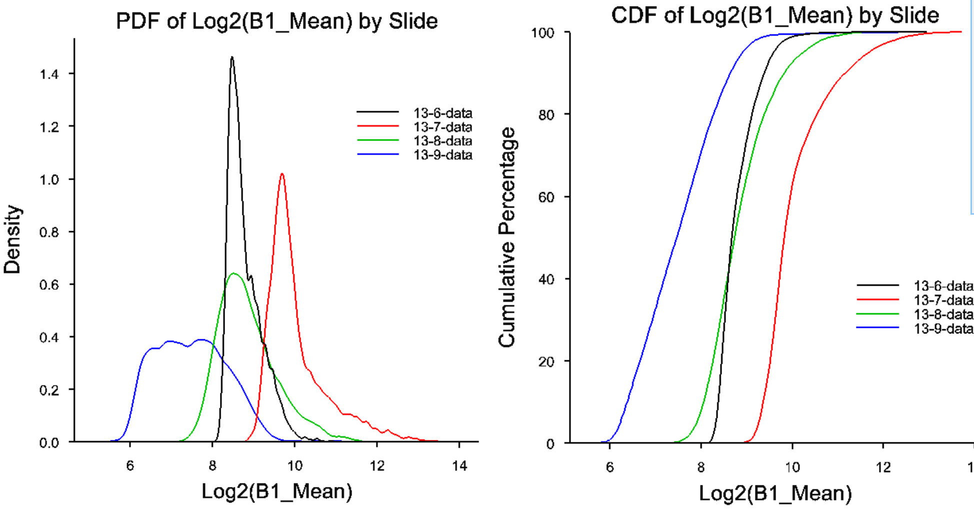


Microarray Density Plot Options

Graphics

☐ Histograms of Intensities by slide
☑ Probability Density of Intensities by slide
☑ Cumulative Density of Intensities by slide
☐ Use Trellis plot for CDFs/PDFs
☑ Display Key

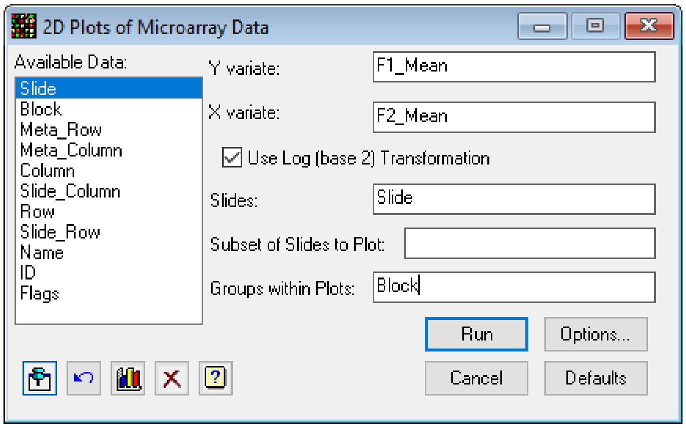
Bandwidth for smoothing PDF: *



The first graph gives the probability density function (*pdf*) of the $\log_2$ transformed red background values as estimated by a kernel density function for each slide. This shows similar trends as the Histogram menu, but as the *pdf* curves can be superimposed on the same graph, the comparisons are easier to make. However, these curves require smoothing of the data, and may miss small-scale features that a histogram may detect. The second curve is cumulative density function (*cdf*), again estimated from a kernel density estimator, and allows estimates of the quantiles of the distributions, so for example the median values of the logged values can be read off by obtaining the points on the curves where the cumulative percentage is 50.

Explore 2d plots

The Density menu allows you to look at the relationship between 2 measurements for each slide. Common plots used for microarray are the M-A plot that plots the log of ratio of red to green against the combined intensity of the two colours. These two statistics can be calculated using the Calculate | Log-ratios menu covered in the next section. For example, using that data from the Histogram example above, and plotting the variate F1_Mean against F2_Mean on the $\log_2$ scale (the mean foreground levels of red and green respectively) with the 2D Plots menu, completed as shown to the right, and with options set as on the next page, gives the following plot. Note, Block indexes the print tips that printed the slide, and so on each slide, the different print tips will be plotted in separate colours, allowing us to look for print tip effects.



2D Plots of Microarray Data

Available Data:

Slide
Block
Meta_Row
Meta_Column
Column
Slide_Column
Row
Slide_Row
Name
ID
Flags

Y variate: F1_Mean

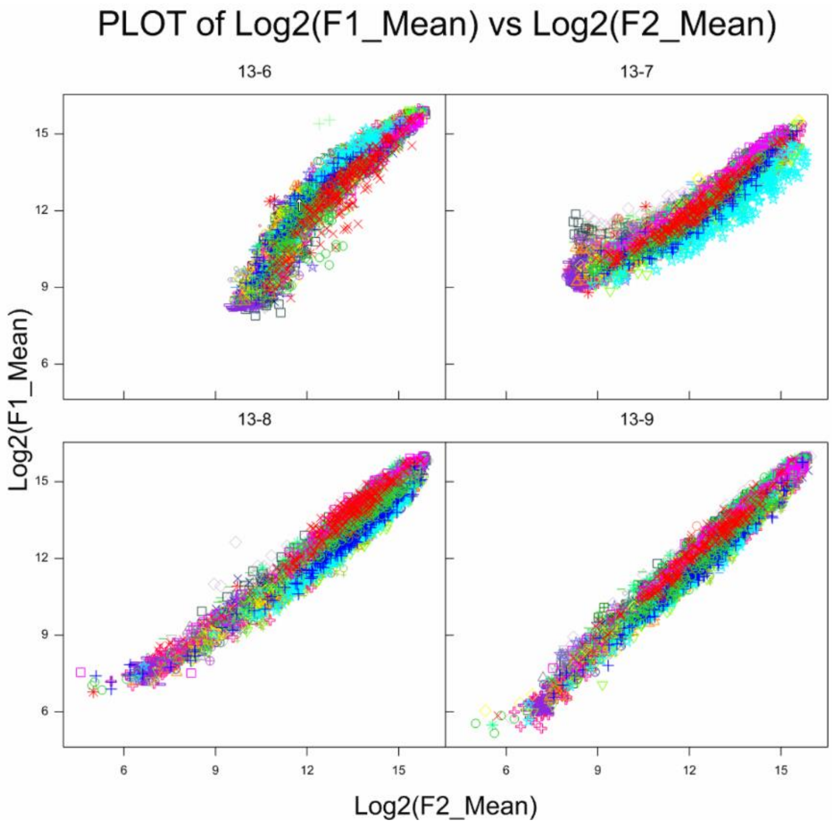
X variate: F2_Mean

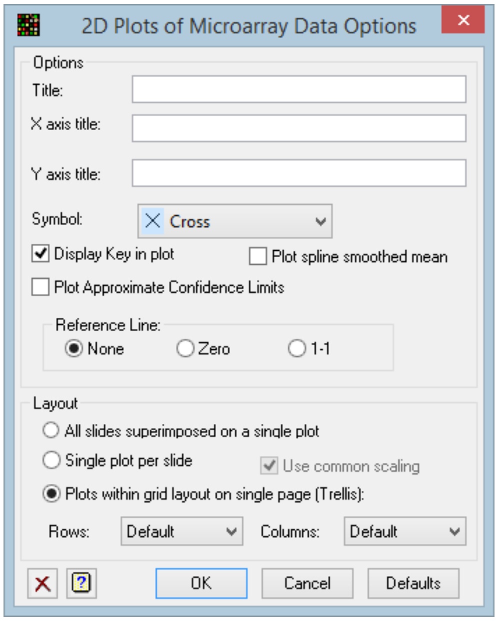
☑ Use Log (base 2) Transformation

Slides: Slide

Subset of Slides to Plot:

Groups within Plots: Block




2D Plots of Microarray Data Options

Options

Title:

X axis title:

Y axis title:

Symbol: Cross

☑ Display Key in plot ☐ Plot spline smoothed mean

☐ Plot Approximate Confidence Limits

Reference Line: ● None ○ Zero ○ 1-1

Layout

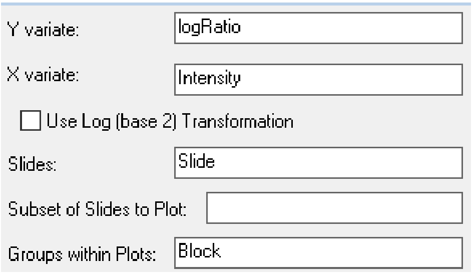
○ All slides superimposed on a single plot

○ Single plot per slide ☑ Use common scaling

● Plots within grid layout on single page (Trellis):

Rows: Default Columns: Default

OK Cancel Defaults


Y variate: logRatio

X variate: Intensity

☐ Use Log (base 2) Transformation

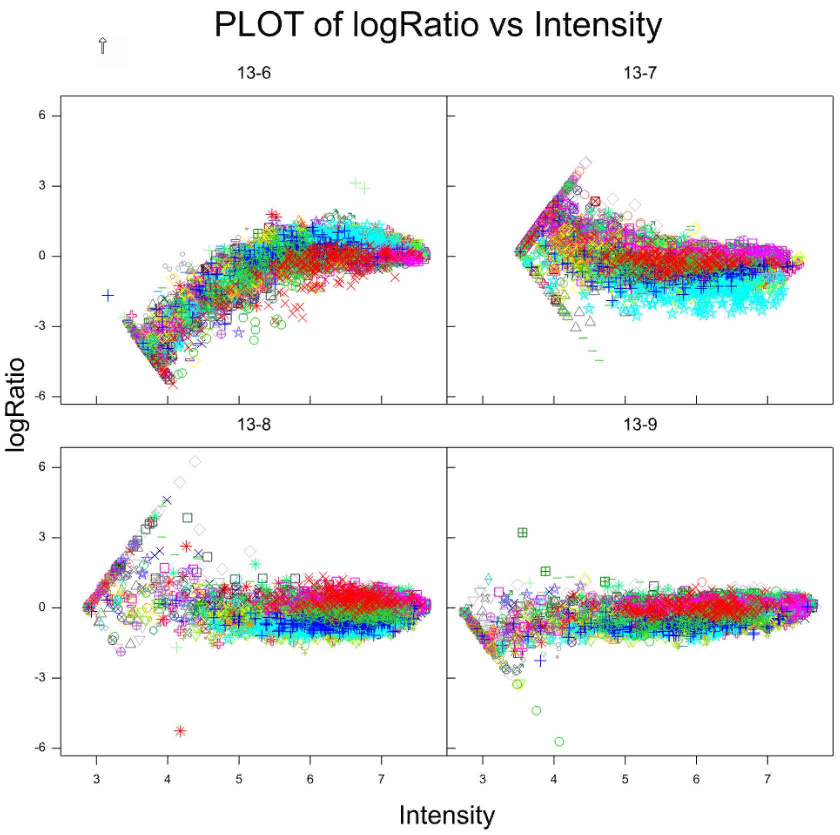
Slides: Slide

Subset of Slides to Plot:

Groups within Plots: Block

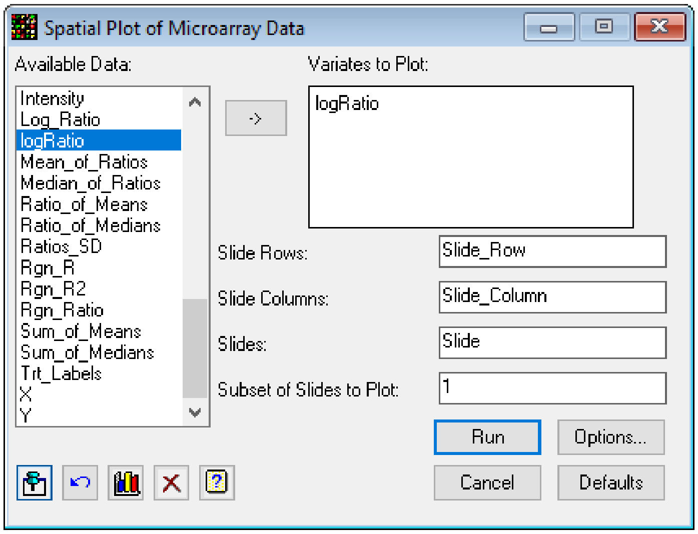
It can be seen that the relationship between red and green is different on the different slides. On slide 13-6, the relationship curves down, whereas on 13-7 it curves up. In addition, the spread around the mean curve on slide 13-8 and 13-9 is less than that on 13-6 and 13-7. The M-A plot is a 45-degree rotation of the red vs. green plot and is obtained by plotting the log-ratio versus the intensity. The M-A plot corresponding to the graphs above is shown below.

The truncation of the points at low levels of intensity comes from limiting the red and green values to be above a minimum value in the Calculate Log-ratios menu. The increased scatter at the low intensity is due to the background levels having been subtracted from the foreground values, giving unstable log-ratios when the foreground and background levels are very similar.



Explore spatial plots

The Spatial plots menu creates a shade plot of the selected variate, with its level being plotted as a coloured cell in a two-way spatial layout, typically using the row and column information from the slide. This allows you to look for trends over the slides, and for areas of the slides where there may be problems due to high backgrounds, scratches, or printing problems (e.g., a pin blocking). The colours representing high and low values can be selected in the options dialog, and the colours used for a given level are interpolated between the two extreme colours chosen. For example, the menu to the right gives a spatial plot of the log-ratio for slide 13-6. The options are set on the Options dialog (right), and here the colour chosen for low values is blue (which represents a higher level of the green dye), and for high values, yellow was chosen (blue and yellow are complementary colours and the scale is visible for colour-blind people). Alternatively, the natural colouring of green for low, and red for high, could be chosen. One option to improve sensitivity around the centre of the distribution is rescaling the log-ratio shading steps from equal steps on the natural scale, to equal steps on the percentile scale. When a measurement follows a normal distribution, most of the colour range will be used in the sparser tails. A rescaling of the data by effectively plotting ranks rather than the raw data, will give more sensitivity around the median and can be achieved by selecting the Percentile option for the Shading Scale. The graph below right uses this scaling and the green/red colours for the plot. With this scaling, you can see the band of green through the centre of the slide more clearly. In the blue/yellow scaling, you can see the lines of blank spots, which are biased toward green more clearly, and the block of greenish spots around row 30, columns 45-70.



Spatial Plot of Microarray Data

Available Data:
- Intensity
- Log_Ratio
- logRatio
- Mean_of_Ratios
- Median_of_Ratios
- Ratio_of_Means
- Ratio_of_Medians
- Ratios_SD
- Rgn_R
- Rgn_R2
- Rgn_Ratio
- Sum_of_Means
- Sum_of_Medians
- Trt_Labels
- X
- Y

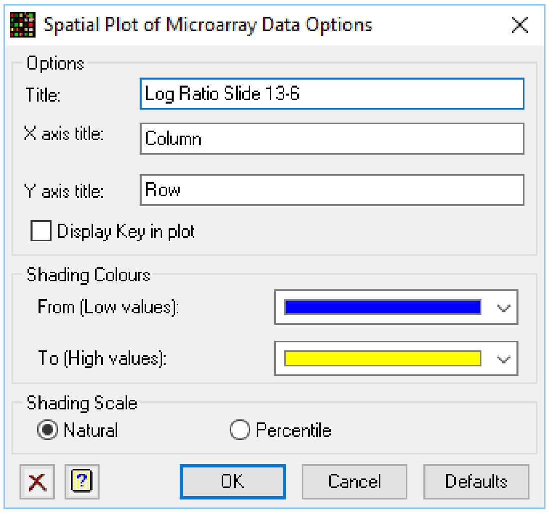
Variates to Plot: logRatio

Slide Rows: Slide_Row

Slide Columns: Slide_Column

Slides: Slide

Subset of Slides to Plot: 1



Spatial Plot of Microarray Data Options

Options

Title: Log Ratio Slide 13-6

X axis title: Column

Y axis title: Row

☐ Display Key in plot

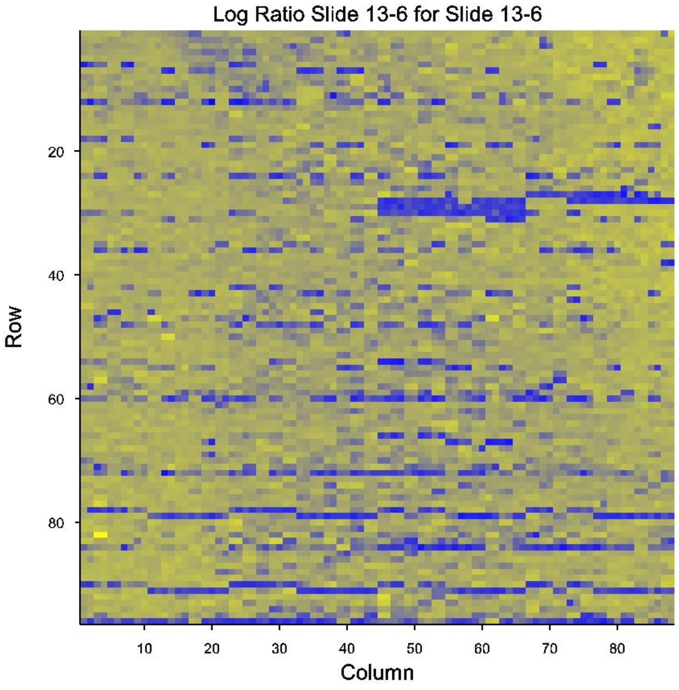
Shading Colours

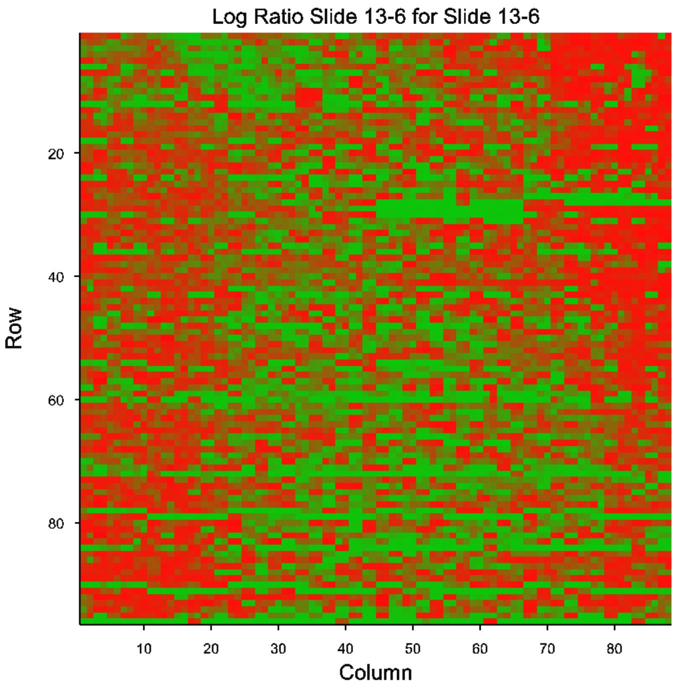
From (Low values): 🟦

To (High values): 🟨

Shading Scale

● Natural ○ Percentile



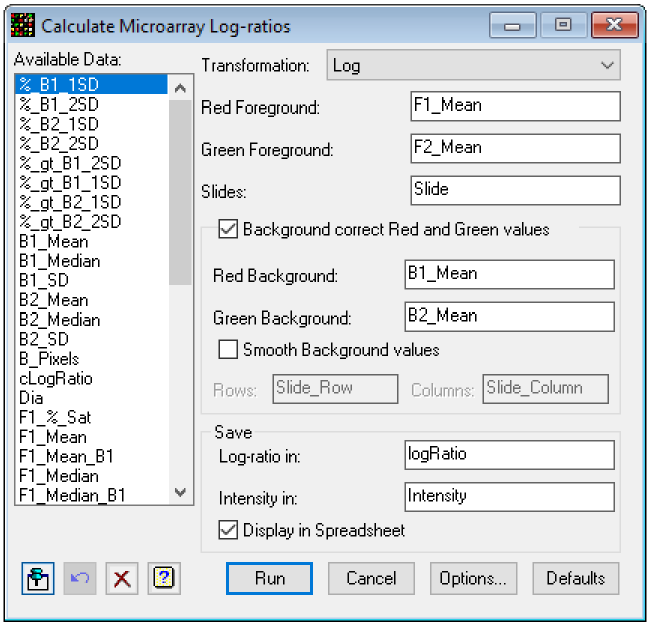


Calculations for microarray data

The Calculate menu allows you to calculate measurements of differential expression for two channel microarrays (log-ratios), or levels of expression for Affymetrix microarrays (Affymetrix expression values). The two types of slides (two colours/single colour) have very different methods of calculating expression levels. Two colour microarrays use relative expression levels, whilst the single colour Affymetrix microarrays calculate an absolute value, averaged over the probes belonging to the gene.

Calculate log-ratios

For a two-colour microarray experiment, we need to calculate the relative level of differential expression between the two targets on the slide. The log-ratio of red to green is the usual measure of differential expression. The base used for the logarithm is usually base 2 so that +1 is equivalent to a double, -1 to half the quantity of red relative to green, and a value 0 indicates equal amounts of the two dyes. The data are log transformed to stabilize the variance of the data. The log-ratio is also equivalent to the difference between the log red and log green values. Another useful statistic, which is independent of the log-ratio, is the intensity (on the log scale) which is calculated as the mean of the log red and log green. In fact, the transform from log red and log green to log-ratio and intensity corresponds to a 45-degree rotation of the variates. Generally, it is found that there is a relationship between the mean and variance of log-ratio and intensity, and this should be corrected for in any analysis. The process of normalization adjusts the log-ratios so that mean relationship with intensity is removed, as there is no reason that low or high abundance spots should be differentially expressed, and this is normally a differential response in the dyes to binding to the probes, or problems with background levels of one of the dyes. The log-ratios and intensity for each spot can be calculated with the Calculate | Log-ratio menu which gives the dialog shown to the right. This calculation has a range of options that will be discussed below.



Transformation: Log

Red Foreground: F1_Mean

Green Foreground: F2_Mean

Slides: Slide

☑ Background correct Red and Green values

Red Background: B1_Mean

Green Background: B2_Mean

☐ Smooth Background values

Rows: Slide_Row Columns: Slide_Column

Save

Log-ratio in: logRatio

Intensity in: Intensity

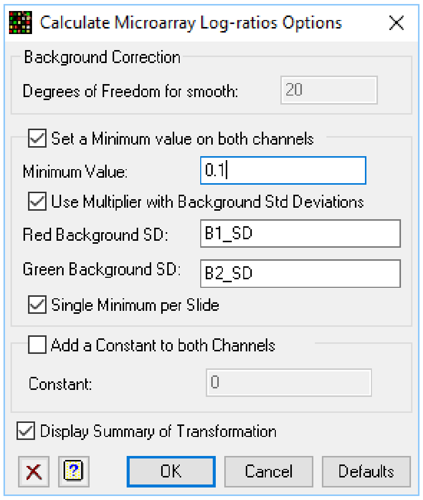
☑ Display in Spreadsheet

Background correction

One approach when calculating log-ratios is to remove the estimated background levels of red and green to allow for trends across the slide in the backgrounds. It is hoped that by doing this, that the log-ratios will be more accurate through the removal of unwanted noise. The background correction will often have the effect of increasing the variance of the log-ratios at the low intensities. The Calculate Log-ratios dialog above right shows this being done for the data introduced in the Histograms menu. One consequence of background correction is that the log-ratios will become undefined where spots may have foreground levels that are below their background, as you cannot take the log of a negative number. Where this happens for both red and green, then there is no valid information on the level of differential expression, and we can insert a missing value for the log-ratio in this circumstance with no loss of information. However, where one channel is above background and the other below, making the log-ratio missing will lose information. In this case, there are options to set a minimum value on each channel. The options are:

1. Set a single minimum value on both colours,
2. Set the minimum value per spot, based on the standard deviation of the background of each colour around the spot,
3. Use an average of the standard deviation of the backgrounds over the whole slide to set the minimum value per colour.

These options can be set by clicking the **Options** button on the menu window. The dialog to the right will appear. To use option 1 above, set "Set a minimum value on both channels" to on (ticked); for option 2, additionally set **Use Multiplier with Background Std Deviations** and give the background standard errors, and for option 3, set **Single Minimum per Slide** to on.


Calculate Microarray Log-ratios Options

Background Correction

Degrees of Freedom for smooth: 20

☑ Set a Minimum value on both channels

Minimum Value: 0.1

☑ Use Multiplier with Background Std Deviations

Red Background SD: B1_SD

Green Background SD: B2_SD

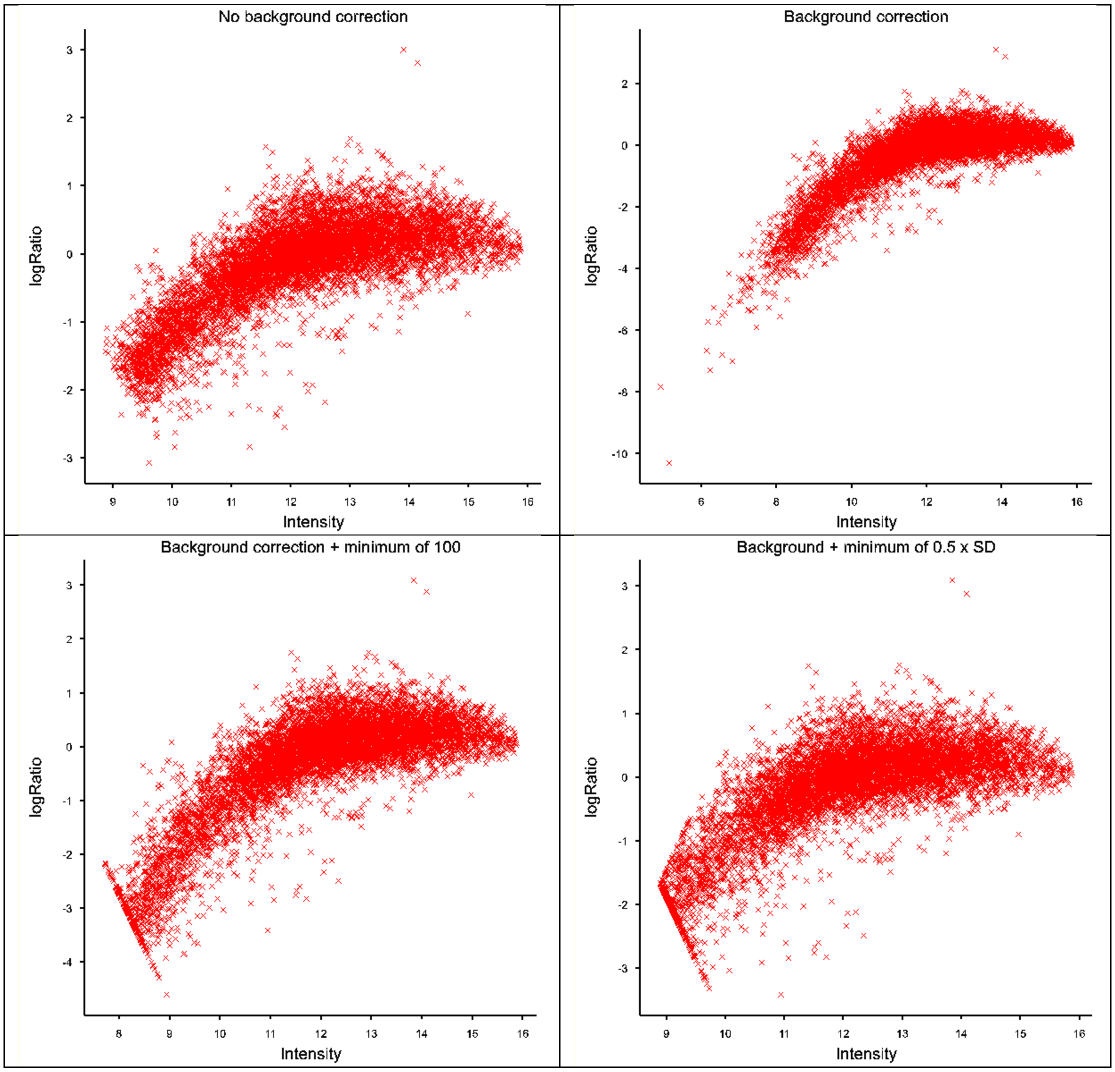
☑ Single Minimum per Slide

☐ Add a Constant to both Channels

Constant: 0

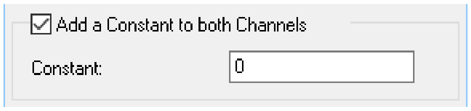
☑ Display Summary of Transformation

To reduce variance for log-ratios at low intensities, you can add a given constant to each channel. The effects some of these options can be seen in the following series of graphs, which are for the slide 13-6 from the earlier example.



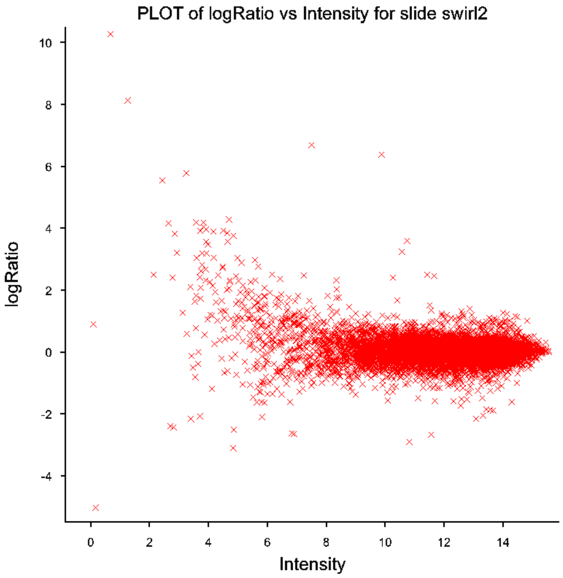
It can be seen that including background correction reduces the scatter from the first to the second graph. Adding in a minimum value for each colour reduces the scatter/tail at the low intensities, and using value based on the standard deviations of the backgrounds means that the minimum value for the green is larger than that for the red.

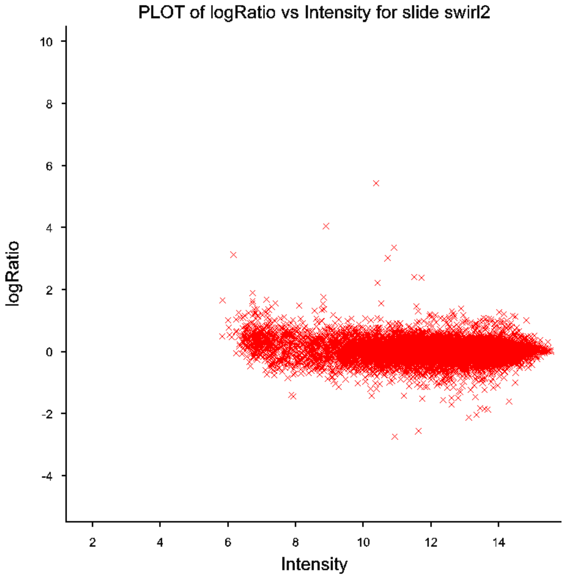
You can also add a constant to both colours to reduce the increase in variance at the lower intensities, by specifying this in the Calculate Log-Ratios Options dialog as shown to the right. The graph below right shows the M-A plot for the data in “Swirl.gsh” for the second slide (swirl2). If you add a constant of 100 to each colour, you obtain the graph to the left, which has roughly constant variance over the range of intensities.



☑ Add a Constant to both Channels

Constant: 0

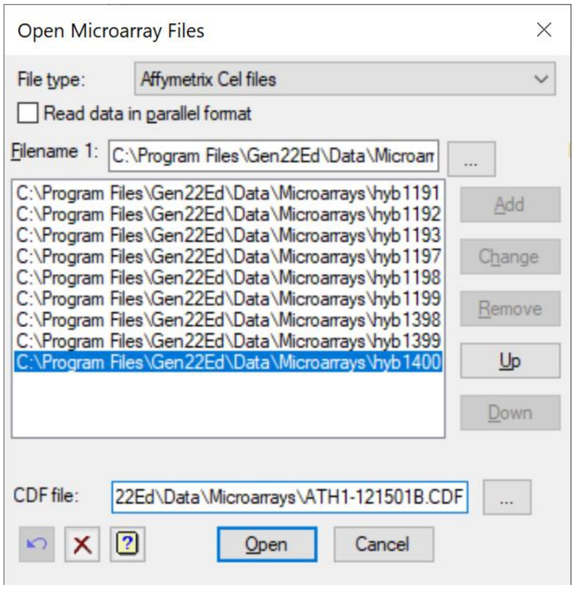

PLOT of logRatio vs Intensity for slide swirl2


PLOT of logRatio vs Intensity for slide swirl2

Once you have calculated the differential expression you are ready to analyse the data. The curvature in some of the previous M-A plots indicate that normalization with respect to the intensity is required. Once you have calculated log-ratios, the next step is then normalization.

Calculate Affymetrix expression values

For Affymetrix data that has been loaded into a spreadsheet, we can use the menu Stats | Microarrays | Calculate | Affymetrix to summarize the results (pairs of PM/MM per gene) to a single expression value per gene. The CEL files “hyb1191.CEL” – “hyb1400.CEL” will be used in the examples in this section. These chips are for the Arabidopis plant and are laid out with a 712 x 712 grid, giving over 500,000 cells per slide. Reading this data into a spreadsheet is very demanding on RAM and requires 300MB. For this reason, this type of data is normally handled through a batch process that only holds one slide in memory at a time. The dialog to the right shows the 9 CEL files being loaded, along with the CDF file for these chips “ATH1-121501B.CDF”. On clicking **Open**, the Affymetrix CEL Read Options dialog appears (as on the following page), which allows the batch processing, which is much faster than creating a spreadsheet. The batch processing options are fewer than provided in the Affymetrix Expression value dialog, which may be a reason not to use batch processing.



Open Microarray Files

File type: Affymetrix Cel files

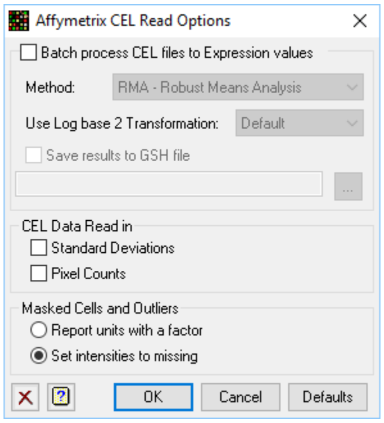
□ Read data in parallel format

Filename 1: C:\Program Files\Gen22Ed\Data\Microarr

C:\Program Files\Gen22Ed\Data\Microarrays\hyb1191
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1192
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1193
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1197
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1198
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1199
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1398
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1399
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1400

CDF file: 22Ed\Data\Microarrays\ATH1-121501B.CDF

There are two optional columns of data that can be read in from a cell file, the standard deviation of the pixels in each cell, and the number of pixels used in calculating the mean of each cell. If you need these columns, select them under the CEL Data Read in options, but not selecting these columns considerably reduces the memory required. The CEL files also contain information on cells that have been masked out as bad cells or detected as outliers by the image analysis software. A factor that holds these flags can be created if the option for **Masked Cells and Outliers** is set to **Report units with a factor**, otherwise the cell intensities will be set to a missing value when a cell is flagged as masked or an outlier (**Set intensities to missing**). On clicking **OK**, we get the spreadsheet below.


Affymetrix CEL Read Options

□ Batch process CEL files to Expression values

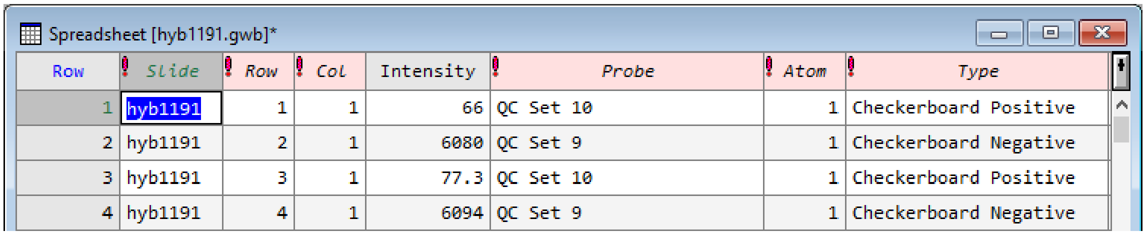
Method: RMA - Robust Means Analysis

Use Log base 2 Transformation: Default

□ Save results to GSH file

CEL Data Read in
□ Standard Deviations
□ Pixel Counts

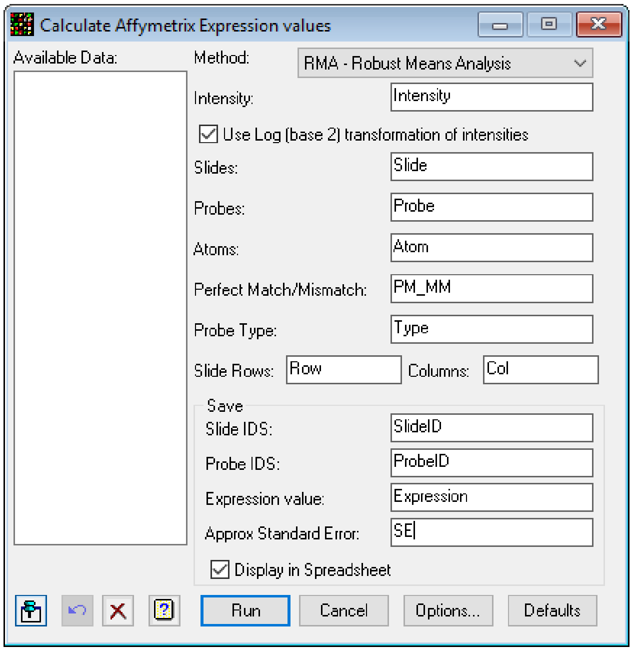
Masked Cells and Outliers
○ Report units with a factor
● Set intensities to missing


Spreadsheet [hyb1191.gwb]*

Row	Slide	Row	Col	Intensity	Probe	Atom	Type
1	hyb1191	1	1	66	QC Set 10	1	Checkerboard Positive
2	hyb1191	2	1	6080	QC Set 9	1	Checkerboard Negative
3	hyb1191	3	1	77.3	QC Set 10	1	Checkerboard Positive
4	hyb1191	4	1	6094	QC Set 9	1	Checkerboard Negative

The column Slide in this spreadsheet indexes the file that the data came from, Row and Col are the spatial position of the cell on the chip and Intensity is the amount of biotin dye bound to the cell (averaged over the pixels in the cell). The Atom factor indexes the pairs of PM/MM cells within a gene, and the factor PM_MM indicates whether the cell is a perfect match (PM) or mismatch (MM) cell. The factor Type gives the type of cell on the chip, with a range of quality control cells on the chips (which are not used currently in the analysis). The cells used to detect genes are of type ‘Expression’, ‘Genotyping’ or ‘CustomSeq’ depending on the type of chip.

Note, with large Affymetrix data sets, if you do not have enough memory, it is better to use the “Spread | Update | Using fast load” menu to update the server. This closes the spreadsheet before reading the data into the server using the SPLOAD directive, which is much faster and saves having the data twice in memory.


Calculate Affymetrix Expression values

Available Data:

Method: RMA - Robust Means Analysis

Intensity: Intensity

☑ Use Log (base 2) transformation of intensities

Slides: Slide

Probes: Probe

Atoms: Atom

Perfect Match/Mismatch: PM_MM

Probe Type: Type

Slide Rows: Row Columns: Col

Save
Slide IDS: SlideID
Probe IDS: ProbeID
Expression value: Expression
Approx Standard Error: SE

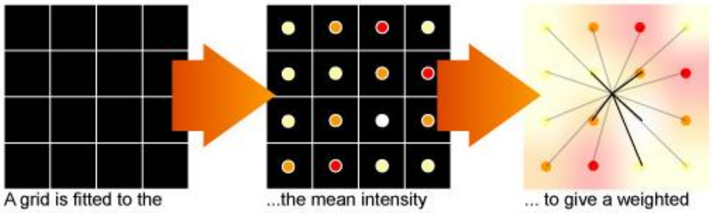
☑ Display in Spreadsheet

The columns are now entered into the Calculate Affymetrix Expression values menu as shown to the above right. There are 3 main algorithms for summarizing the Affymetrix data, RMA (with an alternative algorithm for estimating the parameters called RMA2), MAS4 and MAS5 which are algorithms developed by Affymetrix. MAS4 is an older algorithm and is only provided for completeness and has been superseded by MAS5. These algorithms are explained in the next section.

RMA algorithm

RMA stands for robust means analysis, and it involves 3 steps, background correction - where an error component of the intensities is estimated and eliminated; quantile normalization – where every slide is normalized to have the same cumulative frequency distribution; and summarization – where the median value per probe set, adjusted for slide differences are calculated. This results in an expression value for every gene on every slide. These steps will be explained briefly below.

Background correction

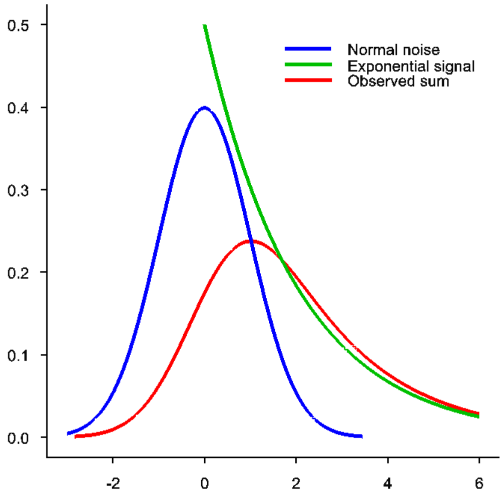


A grid is fitted to the array (the default is 4x4)...

...the mean intensity for the bottom 2% of signal spots for each zone is calculated....

... to give a weighted sum of the zone backgrounds. This is subtracted from each spot.

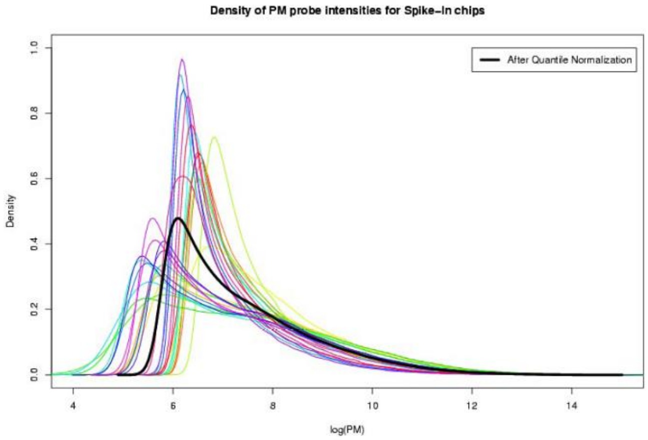
The option of removing a background value allowing for trends across the slide is available with all algorithms. This uses the mean of the lowest 2% of cells in 16 zones (in a 4 x 4 layout) over the slide, and then forms the weighted average of these for each cell, with the weights depending on the squared distance from the zone centroids to the cell. There are two weighting schemes available as options: the Affymetrix weighting which adds a smoothing constant to the denominator of the weights (i.e. $1/(d^2+S)$); and Distance weighting which smooths by not letting the weights go above a certain minimum constant (i.e. $1/(\min(d^2, S))$. The alternative weighting to the Affymetrix standard is provided since the addition of the constant induces strange effects around the edges of the chip, particularly in the corners.



The next background correction step in RMA is to fit a noise model to the intensities from the PM cells. The RMA analysis ignores the MM cells as it assumes that these contain gene information, rather than just cross hybridization levels, and so removing these reduces the signal in the gene expression data. The PM intensities are assumed to come from the sum of a normal distribution (noise) and an exponential distribution (signal). Thus, if $\boldsymbol{z}$ is the observed intensity, then $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$, where $\boldsymbol{x} \sim N(\mu, \sigma)$ and $\boldsymbol{y} \sim \mathrm{Exp}(\alpha)$. The parameters, μ, σ and α, are estimated by maximum likelihood estimators (which can have problems with convergence in some cases). The RMA2 algorithm just uses different estimators for the parameters μ, σ and α based on moments, which is faster and does not have convergence problems.

Normalization

The next step in the RMA analysis is normalizing the background corrected results over the slides so that each slide's results have the same cumulative density (quantile normalization). This seems an extreme normalization but is performed so that the levels of differential expression on each slide have the same profile. It is justified from the application of the technique to a few studies with known outcomes giving results that are more accurate. The algorithm for this is to sort the intensities on each slide, average the results at each rank over the slides (using a median, mean, or geometric mean), and then replace the values on all slides with the averages. The graph above shows a set of slides to be quantile normalized. All the individual slides will be normalized to the black curve that is the average profile.



Summary over Probes

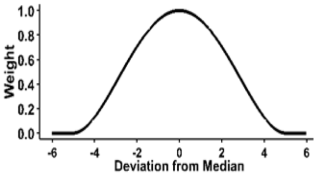
The final step in the RMA analysis is to create an average of the 11-20 probes representing a gene. The algorithm used for this is to take the medians over the probes and adjust for any overall slide effects. The joint estimation of the probe and slide medians requires an iterative algorithm that switches between estimating the probe and slide medians, removing these from the results sequentially until convergence. The use of medians is regarded as providing robustness, which gives the algorithm its name.

MAS 4.0 algorithm

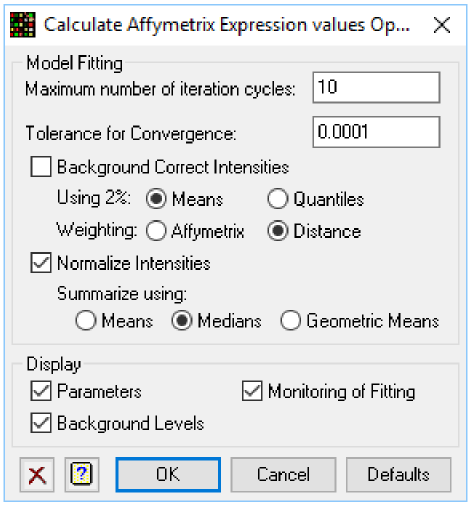
The MAS 4.0 algorithm uses the same background correction as described in the RMA algorithm, where the lowest 2% of intensities are used to adjust the data. The PM values are then corrected for cross-hybridization by subtracting the corresponding MM value. The resulting intensities for each gene are then averaged over the probes by using a winsorized average, where the minimum and maximum intensities are eliminated along with any intensity greater than 3 standard deviations from the mean. The final restriction using 3 standard deviations from the mean is redundant unless a gene has more than 13 probes.

MAS 5.0 algorithm

The MAS 5.0 algorithm uses the same background correction as described previously. The PM values are then corrected for cross-hybridization by subtracting an ideal MM value, which is calculated using the values of all the probes. The ideal mismatch is the actual MM if it is less than PM. The robust average of the MM for the gene is used if this is less than the PM; otherwise, a value just less than the PM is used. The resulting intensities for each gene are then averaged over the probes by using a Tukey-Biweight average, where the intensities are weighted according to their distance from the median. If s is the median absolute deviance from the median, m, then the weights are 0 if $|x - m| > 5$, and $(1 – (|x - m|/5s)^2)^2$ otherwise (plotted to the right).



The algorithm to use is selected from the dropdown list in the Calculate Affymetrix Expression values menu. The options for this menu (shown right) control the iteration cycles and convergence criterion (tolerance) for the median calculations in the RMA analysis, the background correction used in all analyses, and the quantile normalization used in the RMA analysis. This dialog also controls what output is displayed in the Output window.



Calculate Affymetrix Expression values Op...

Model Fitting

Maximum number of iteration cycles: 10

Tolerance for Convergence: 0.0001

☐ Background Correct Intensities

Using 2%: ● Means ○ Quantiles

Weighting: ○ Affymetrix ● Distance

☑ Normalize Intensities

Summarize using:

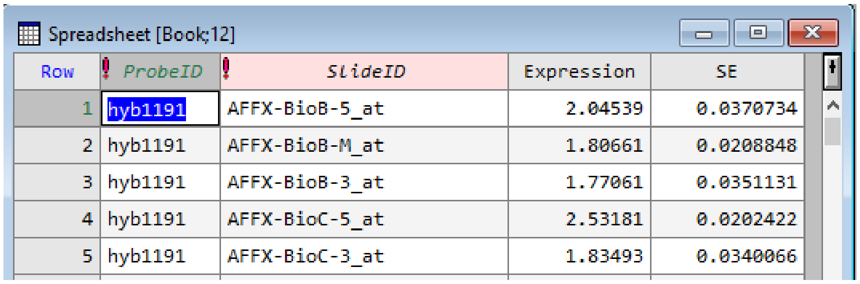
○ Means ● Medians ○ Geometric Means

Display

☑ Parameters ☑ Monitoring of Fitting

☑ Background Levels

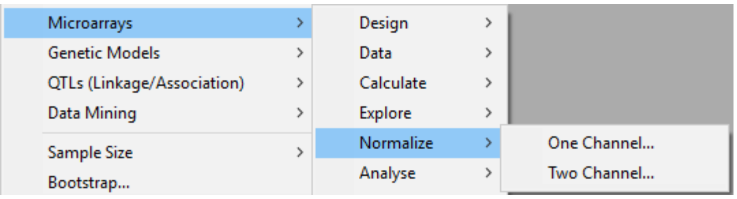
Performing an RMA analysis on the Arabidopis dataset read in previously took 55 minutes CPU time on a 2 GHz Pentium 4 PC. The results are displayed in a spreadsheet as bellow.



Spreadsheet [Book;12]

Row	ProbeID	SlideID	Expression	SE
1	hyb1191	AFFX-BioB-5_at	2.04539	0.0370734
2	hyb1191	AFFX-BioB-M_at	1.80661	0.0208848
3	hyb1191	AFFX-BioB-3_at	1.77061	0.0351131
4	hyb1191	AFFX-BioC-5_at	2.53181	0.0202422
5	hyb1191	AFFX-BioC-3_at	1.83493	0.0340066

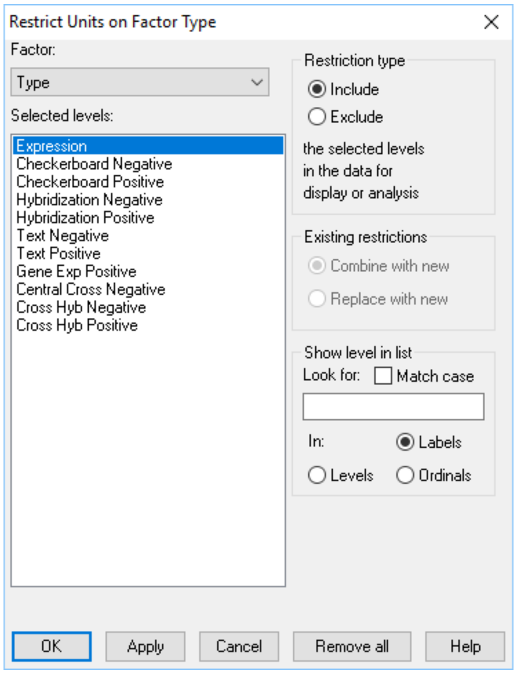
Normalization of microarray data



One channel (Affymetrix)

The quantile normalization, which is part of the RMA analysis described previously, can be applied to any single variable over a set of slides. This menu allows you to apply the quantile normalization to any data set.

The following dataset in “Hyb-PM_MM.gwb” was obtained by restricting the “Hyb-AllData.gwb” file to Type equal to Expression using the Spread | Restrict/Filter | To groups menu (as shown to the right) and then unstacking the data using the Spread | Manipulate | Unstack menu. We put the PM and MM values into two columns by specifying that we unstack Intensity using the factor PM_MM, with the ID factors Slide, Probe and Atom. We drop the ROW, COL, and Type factors as these are now not required. The completed Unstack menu is shown below.


Restrict Units on Factor Type

Factor: Type

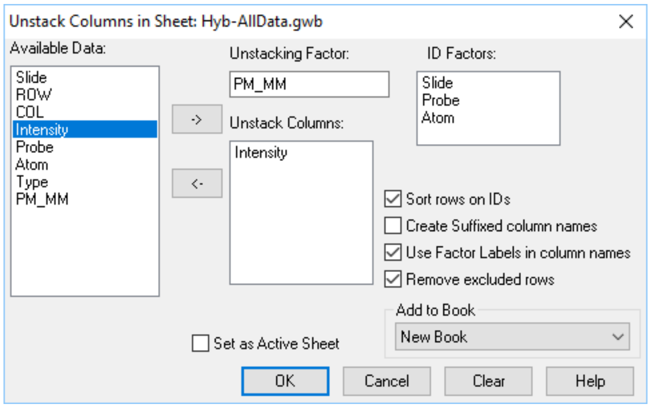
Selected levels:
Expression
Checkerboard Negative
Checkerboard Positive
Hybridization Negative
Hybridization Positive
Text Negative
Text Positive
Gene Exp Positive
Central Cross Negative
Cross Hyb Negative
Cross Hyb Positive

Restriction type
● Include
○ Exclude
the selected levels in the data for display or analysis

Existing restrictions
○ Combine with new
○ Replace with new

Show level in list
Look for: □ Match case

In: ● Labels ○ Levels ○ Ordinals


Unstack Columns in Sheet: Hyb-AllData.gwb

Available Data:
Slide
ROW
COL
Intensity
Probe
Atom
Type
PM_MM

Unstacking Factor: PM_MM

ID Factors:
Slide
Probe
Atom

Unstack Columns:
Intensity

☑ Sort rows on IDs
□ Create Suffixed column names
☑ Use Factor Labels in column names
☑ Remove excluded rows

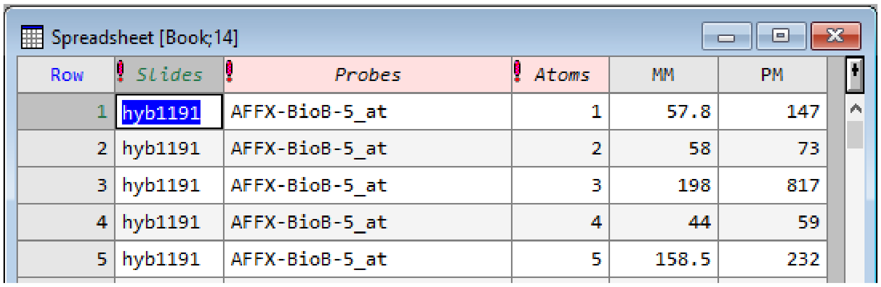
Add to Book: New Book

□ Set as Active Sheet

However, using the Spread menu is very slow compared with creating this spreadsheet with a set of commands. The following short program will create the same spreadsheet in much less time.

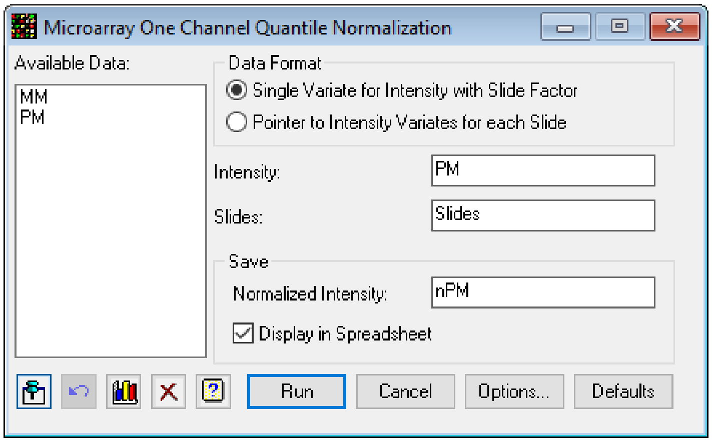
```
SORT    [INDEX=Slide,Probe,Atom] \
        Slide,Probe,Atom,Type,PM_MM,Intensity
SUBSET [PM_MM .in. 'MM' .and. Type .in. 'Expression'] \
        Intensity; MM
SUBSET [PM_MM .in. 'PM' .and. Type .in. 'Expression'] \
        Slide,Probe,Atom, Intensity; Slides,Probes,Atoms,PM
FSPREAD Slides,Probes,Atoms,MM,PM
```

This will generate the spreadsheet to the right (saved as “Hyb-PM_MM.gwb”). The quantile normalization of the PM values in this spreadsheet can be done as follows.


Spreadsheet [Book;14]

Row	Slides	Probes	Atoms	MM	PM
1	hyb1191	AFFX-BioB-5_at	1	57.8	147
2	hyb1191	AFFX-BioB-5_at	2	58	73
3	hyb1191	AFFX-BioB-5_at	3	198	817
4	hyb1191	AFFX-BioB-5_at	4	44	59
5	hyb1191	AFFX-BioB-5_at	5	158.5	232

The Normalize One Channel menu is completed as to the right. In the options we select Geometric Means as the Summary Method, as below.



Microarray One Channel Quantile Normalization

Available Data:
MM
PM

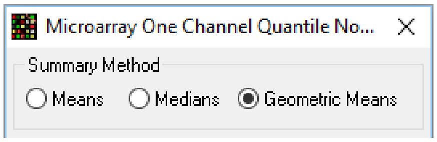
Data Format
● Single Variate for Intensity with Slide Factor
○ Pointer to Intensity Variates for each Slide

Intensity: PM

Slides: Slides

Save
Normalized Intensity: nPM
☑ Display in Spreadsheet

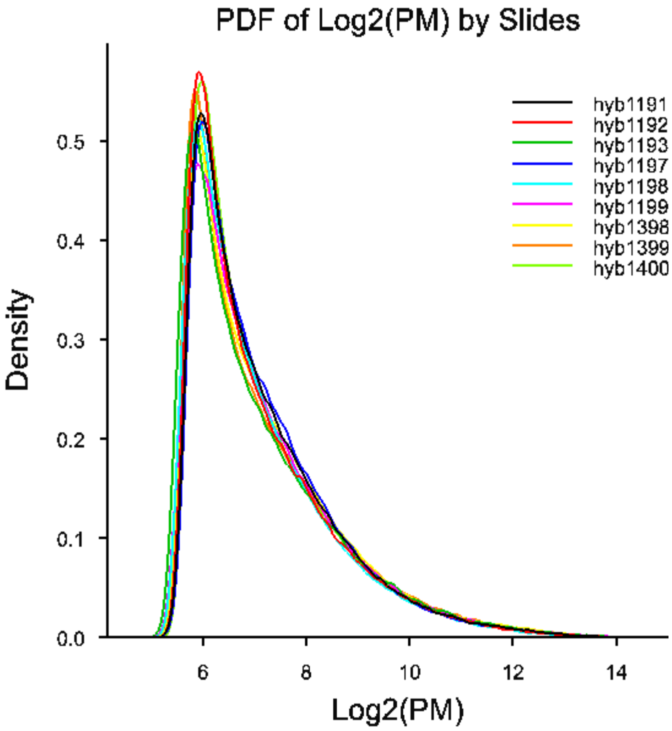
Clicking the **Run** button will now add a new column nPM to the spreadsheet.

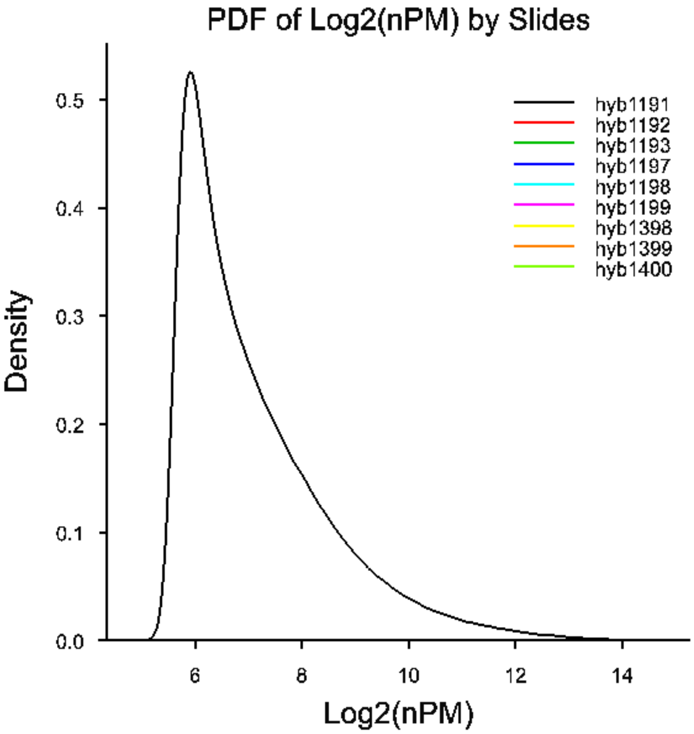


Microarray One Channel Quantile No...

Summary Method
○ Means ○ Medians ● Geometric Means

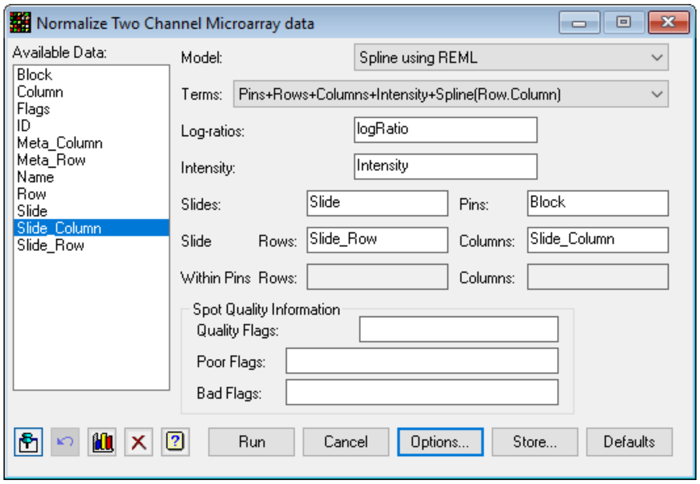
If we had looked at the density of the PM values using the Explore | Density menu we would have discovered that the 9 slides already had very similar density functions, and that a log-transform seems to be required. The log-transformed density plot is shown below left. If we plotted the same density plot of the quantile normalized data, all the density or cdf curves would conicide as below right.





Two channels

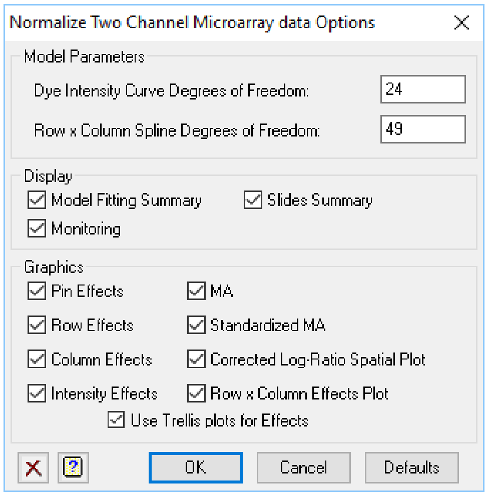
This menu allows the log-ratios from a two-colour experiment to be normalized to remove spatial trends, and dye intensity effects. Removing noise from various sources should improve the ability to detect differential expression for particular genes. The menu is shown to the left. There are two main model-fitting approaches, using REML with splines (Baird *et al.* 2004) or FIT with Loess (Yang *et al.* 2002). The REML model is the recommended model as it is more flexible and can fit a wider range of terms, but the Loess model is provided if the user wants to fit the standard model used in the Bioconductor package of R.
Once the model fitting approach has been decided, there is a drop-down list of models that can be fitted, sorted roughly in order of complexity. The terms that should be fitted are best chosen by examination of the plots from the explore menu, but normally at least the Pins, Rows, Columns, and Intensity effects would be fitted.



Available Data:
Block
Column
Flags
ID
Meta_Column
Meta_Row
Name
Row
Slide
Slide_Column
Slide_Row

Model: Spline using REML
Terms: Pins+Rows+Columns+Intensity+Spline(Row.Column)
Log-ratios: logRatio
Intensity: Intensity
Slides: Slide Pins: Block
Slide Rows: Slide_Row Columns: Slide_Column
Within Pins Rows: Columns:
Spot Quality Information
Quality Flags:
Poor Flags:
Bad Flags:

The Normalize Two Channel Microarray data menu is shown to the right using the data in “Data13-6-9.gwb” examined previously. The model to be fitted is chosen as the most complex one allowing a smooth 2-dimensional surface over the rows and columns, along with pin, row, column, and intensity effects. The only terms not in this model are the AR1 autocorrelation effects and the Intensity x Pin effects. These could be added if they gave a significant increase in the variation explained. The menu is completed as to the above right, and the options are completed in the dialog to the right. The model parameters control the flexibility of the curves to fit to the data. If they are too large, it will considerably increase both the time to fit the model and possibly the variance of the corrected log-ratios. The other options control the output and graphs plotted by the analysis.

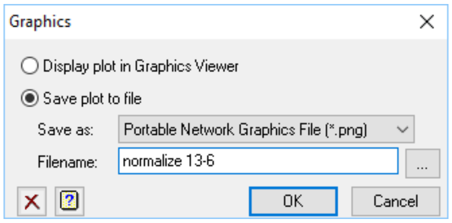


Model Parameters
Dye Intensity Curve Degrees of Freedom: 24
Row x Column Spline Degrees of Freedom: 49

Display
☑ Model Fitting Summary ☑ Slides Summary
☑ Monitoring

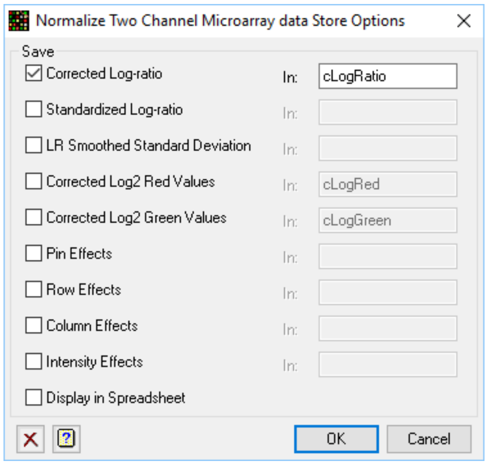
Graphics
☑ Pin Effects ☑ MA
☑ Row Effects ☑ Standardized MA
☑ Column Effects ☑ Corrected Log-Ratio Spatial Plot
☑ Intensity Effects ☑ Row x Column Effects Plot
☑ Use Trellis plots for Effects

There is a graphics button on the main menu, which opens a window (right) that allows the graphics output to be save directly to files for subsequent output, viewing or inclusion into reports.

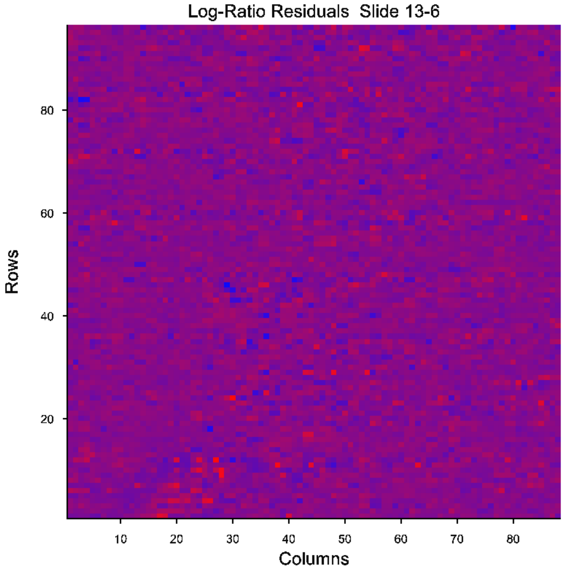


○ Display plot in Graphics Viewer
● Save plot to file
Save as: Portable Network Graphics File (*.png)
Filename: normalize 13-6

The **Store** button on the menu allows the results from the analysis to be saved, and generally, the corrected log-ratio would always be saved, with the other columns being optional. The standardized log-ratio is the corrected log-ratio adjusted for unequal variance over the intensity range. The M-A plot produced in the options menu (above right) will indicate whether the variance does change with intensity, and the standardized M-A plot shows the effect of standardizing for the variation over the range of intensities. The output from the analysis is shown below right and on the next pages.

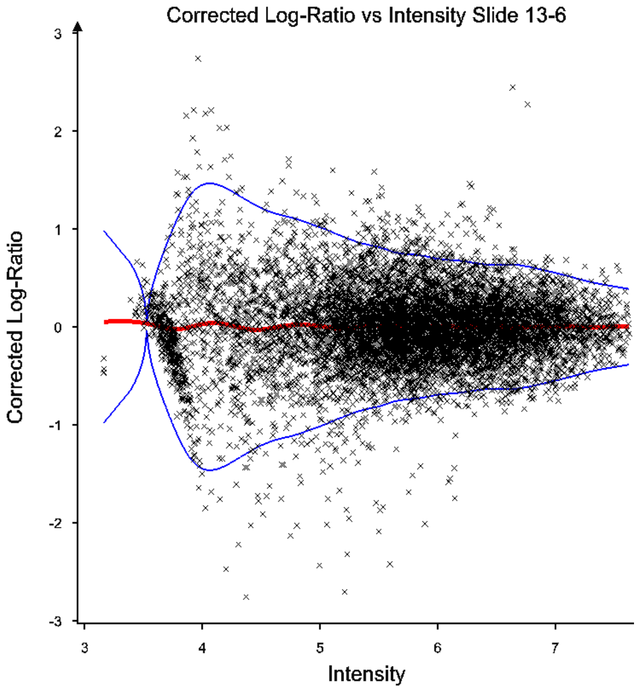


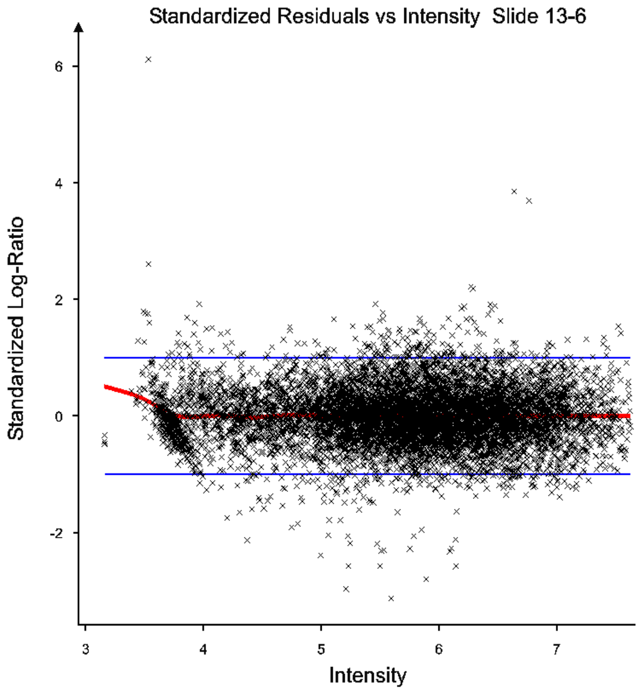
Save
☑ Corrected Log-ratio In: cLogRatio
□ Standardized Log-ratio In:
□ LR Smoothed Standard Deviation In:
□ Corrected Log2 Red Values In: cLogRed
□ Corrected Log2 Green Values In: cLogGreen
□ Pin Effects In:
□ Row Effects In:
□ Column Effects In:
□ Intensity Effects In:
□ Display in Spreadsheet

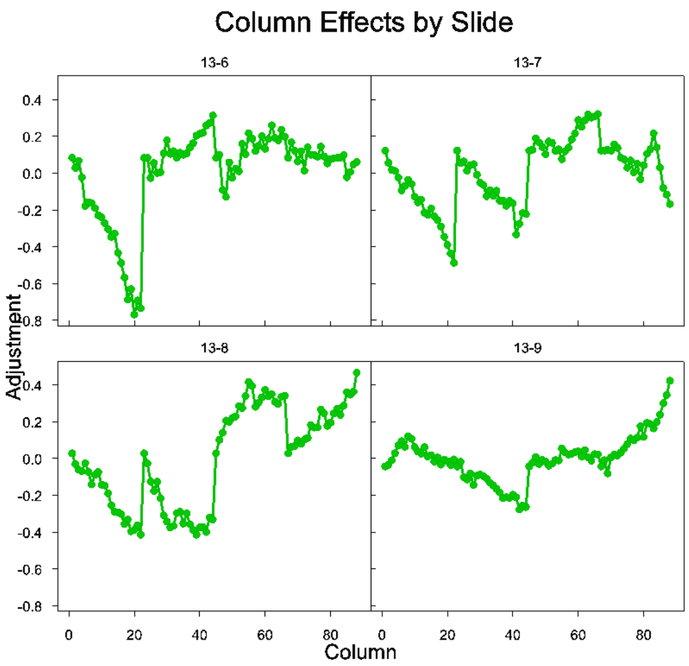


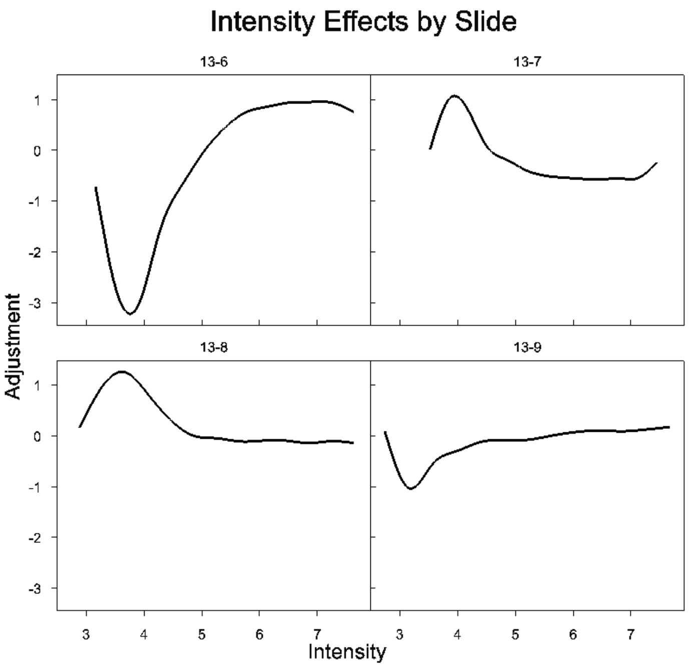
The spatial plots of the residuals for each slide (the one for slide 13-6 only is shown to the right) allow you to check for spatial effects that have not been considered. The plot for 13-6 looks quite clean, but there is perhaps an area at the bottom around column 20 which has too many bright red spots.

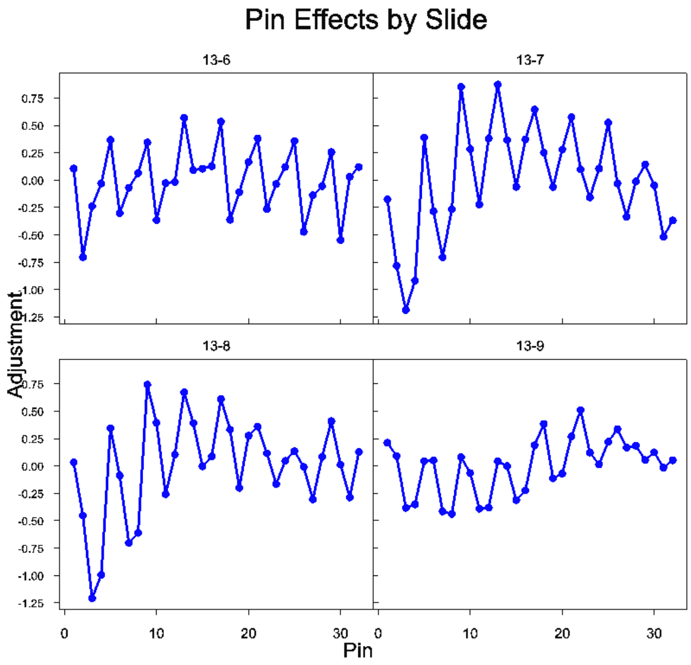
The M-A plot for 13-6 shows that there is a variance changes with intensity. The blue lines give approximate 95% confidence curves for the points, and the red line is the smoothed mean, which runs along at approximately zero as desired. The standardized M-A plot (below right) has divided each log-ratio by the estimated confidence limits used in plotting the blue lines in the plot below left. Thus 95% of the points should lie within +/- 1. If we used the log-ratios from the first plot (below left), there would be a tendency to select points with low intensity as these have the largest variance, and so will produce extreme values more often. Using standardized values would tend to avoid this problem, although in this case the behaviour of the confidence curve around the low intensities would be cause for concern.

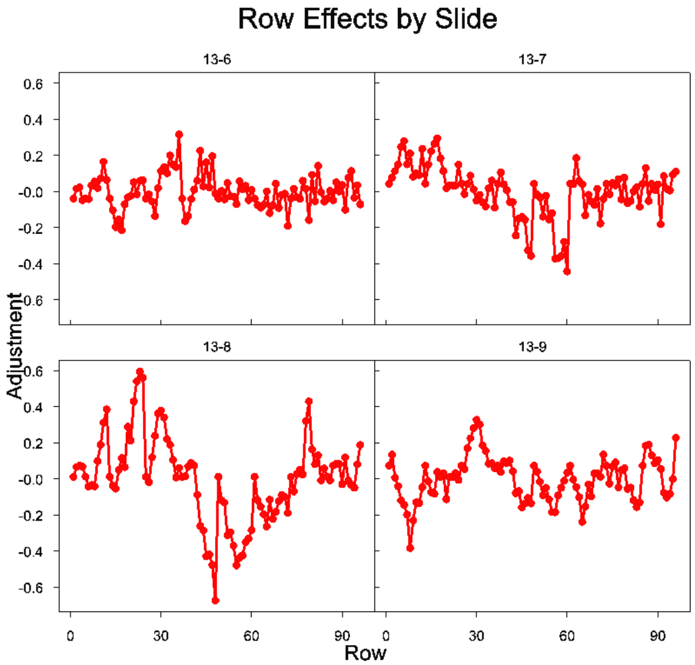






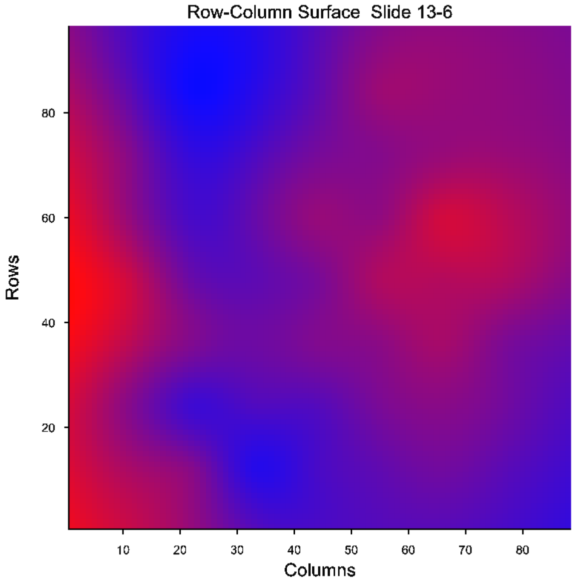




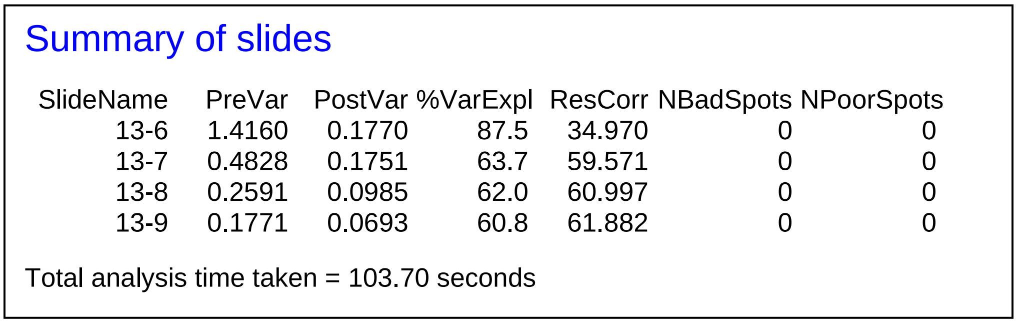


The four graphs above show the estimated pin, row, column, and intensity effects removed across the 4 slides (all plotted in one graph as the Trellis option was selected in Options). They show quite different patterns across the 4 slides, although the alternating pin effects show consistency across the slides. As there are 4 pins across each slide, the alternation in the pin effects is a spatial effect in the columns (every 4th pin block belongs in the same meta-column).

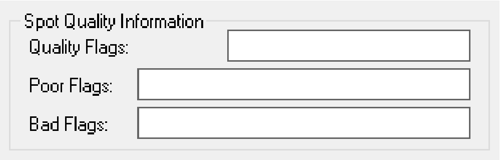
The row x columns effects plot is shown to the right. This graph is a shade plot of the fitted 2d spline for rows by columns effects. This removes the area of low log-ratios through the centre of the slide detected in the graph generated with the Explore | Spatial plot menu.



The output contains the following summary:



```
Summary of slides

 SlideName     PreVar    PostVar %VarExpl  ResCorr NBadSpots NPoorSpots
      13-6     1.4160     0.1770     87.5   34.970         0          0
      13-7     0.4828     0.1751     63.7   59.571         0          0
      13-8     0.2591     0.0985     62.0   60.997         0          0
      13-9     0.1771     0.0693     60.8   61.882         0          0

Total analysis time taken = 103.70 seconds
```

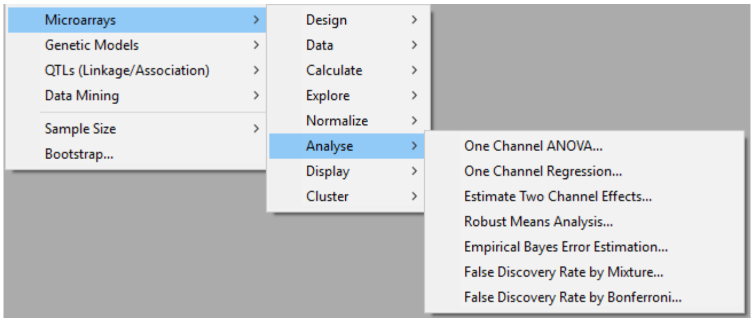


The PreVar column gives the variance of the log-ratio before normalization, and the PostVar column gives it afterwards. The %VarExpl column gives the percentage of variance explained by the model and ResCorr gives the correlation between the adjusted log-ratios and the raw log-ratios. NBadSpots and NPoorSpots give the number of bad and poor spots on the slide as specified by the Spot Quality Information on the Normalize 2 Channel Microarray menu (shown to the right). The quality flags indicate spots that the scanning software has marked as either poor or bad. For example, GenePix uses values of -25 and -50 to indicate poor spots and -75 and -100 to for bad spots.

The increase in the % variance explained is a rough guide as to whether extra terms in the model are explaining more noise. If this does not increase when extra terms are added, then it is not worth adding that model term to the model.

Once you have the normalized microarray data, you are now ready carry out the analysis to get the level of differential expression of the targets for the probes on the slides.

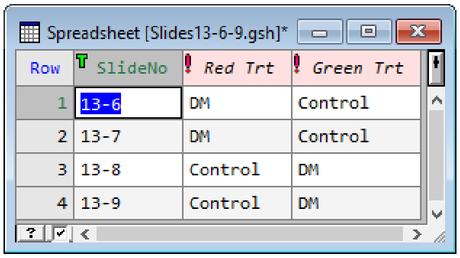
Analysis of microarray data


Microarrays
Genetic Models
QTLs (Linkage/Association)
Data Mining
Sample Size
Bootstrap...
Design
Data
Calculate
Explore
Normalize
Analyse
Display
Cluster
One Channel ANOVA...
One Channel Regression...
Estimate Two Channel Effects...
Robust Means Analysis...
Empirical Bayes Error Estimation...
False Discovery Rate by Mixture...
False Discovery Rate by Bonferroni...

The Stats | Microarrays | Analyse menu provides various menus to calculate summary statistics over the slides in an experiment. The appropriate analysis for the normalized data will depend on whether you are using a single colour or two-colour microarray. For two-colour microarrays, you will need to take the relative differences between targets on the slides and obtain the best estimates of all these over all the slides using the Estimate Two Channel Effects. For single colour microarrays, you will summarize over the replicates using either the Single Channel ANOVA that uses means to summarize the data or the Robust Means Analysis that uses medians to summarize the data.

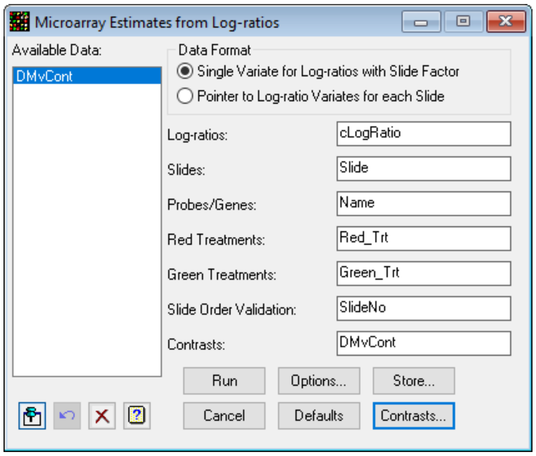
Estimate two-channel effects

This menu summarizes the log-ratios over a series of slides. Apart from the normalized log-ratios for the slides, you will need a spreadsheet giving the targets (treatments) allocated to the red and green dyes on each slide. The spreadsheet requires three columns, the green target/treatment as a factor, the red target/treatment as a factor and the name of the slide as a text or variate to match the slide labels or levels in the log-ratio dataset.


Spreadsheet [Slides13-6-9.gsh]*

Row	SlideNo	Red Trt	Green Trt
1	13-6	DM	Control
2	13-7	DM	Control
3	13-8	Control	DM
4	13-9	Control	DM

For example, for the 4 slides in the experiment in the “Data13-6-9.gwb” spreadsheet, the target spreadsheet is shown to the right. This experiment has just two targets DM and Control, with two replicates of two dye swaps. The data are available in the file “Slides13-6-9.gsh” shown to the above right. This information can be entered into the Estimate Two Channel Effects menu shown to the right (the dialog titled “Microarray Estimates from Log-ratios”).

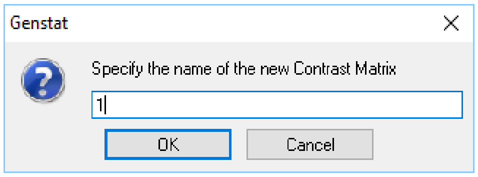

Microarray Estimates from Log-ratios

Available Data: DMvCont

Data Format
● Single Variate for Log-ratios with Slide Factor
○ Pointer to Log-ratio Variates for each Slide

Log-ratios: cLogRatio
Slides: Slide
Probes/Genes: Name
Red Treatments: Red_Trt
Green Treatments: Green_Trt
Slide Order Validation: SlideNo
Contrasts: DMvCont

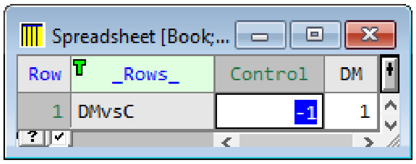
The corrected log-ratio saved from the Normalize menu is entered, as well as the factors that index the slides and probes for this variate. The target information on the red and green dyes is entered from the spreadsheet above. Slide validation order given in SlideNo is not required but is given to check that the order and names of the slides in the target spreadsheet match that in the log-ratios spreadsheet. If the labels do not match between the Slides factor and the Slide Order Validation entries, this will be flagged as an error in the output. If the order between the labels and the Slide Order Validation entries do not match, then a warning will be printed in the output and the Red and Green Treatment factors will be sorted into the order that matches with the Slides factor. If you want to produce some summaries over the treatments (e.g., main effects over factorial combinations or particular comparisons between targets), then you can set up a contrast matrix to specify these. To create this matrix, enter the matrix name in the **Contrasts** field, and click the **Contrasts** button to create the matrix. Give the number of contrasts in the dialog that appears (right), and a matrix will be created for you with the columns headed with the target labels.


Genstat

Specify the name of the new Contrast Matrix

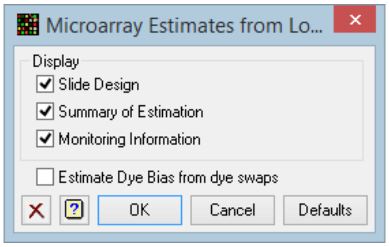
1

In this experiment, which has only two treatments, we only need the difference between the two treatments, and so this contrast of DM – Control is entered as -1 and 1 (right).


Spreadsheet [Book;...

Row	_Rows_	Control	DM
1	DMvsC	-1	1

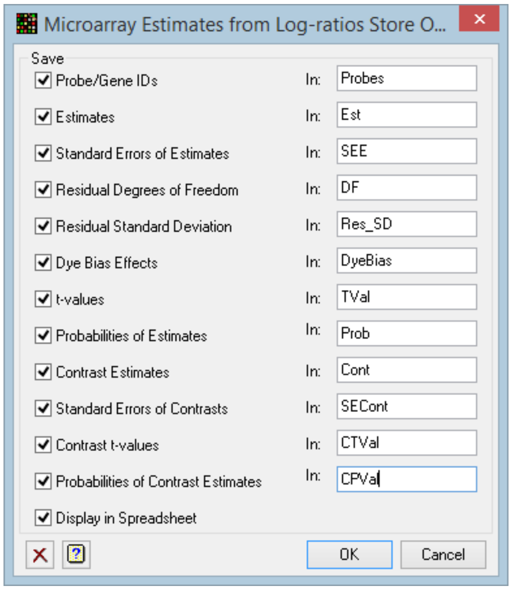
The options menu can now be completed (right). The most important option in this menu is the one that specifies whether a dye bias is estimated from the dye swaps. If there are no dye swaps in the experiment, this must be turned off, as otherwise the treatment estimates will be confounded with the dye effect, and you will get an error in the output that will say that the design is disconnected. If you have balanced dye swaps in your experiment, the means will be the same whether this is selected or not, but the variance of the dye effect will be removed from the residual variation, and one degree of freedom will be taken of the residual degrees of freedom to account for this. In a reasonable sized experiment or in one not balanced for dyes, it is best to estimate the dye bias effect to get an unbiased estimates of effects and their standard errors. In this experiment with only 4 slides, the option to estimate dye bias will not be used to save a degree of freedom in the statistical tests, which decreases the size of effect needed to be significant at the risk of an effect being upset by some dye bias.


Microarray Estimates from Lo...

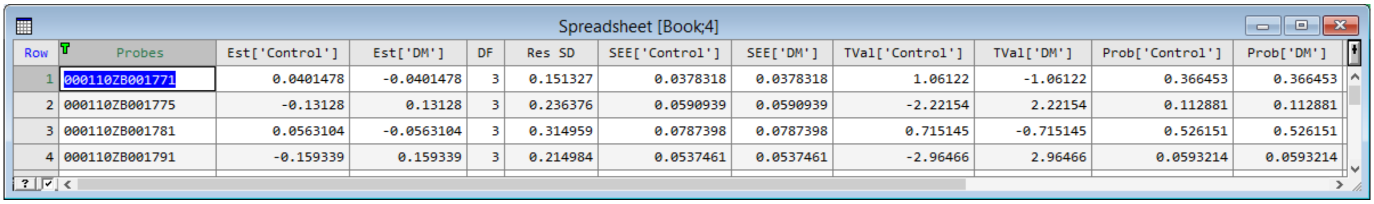
Display
- ☑ Slide Design
- ☑ Summary of Estimation
- ☑ Monitoring Information

□ Estimate Dye Bias from dye swaps

The Store button allows the results to be saved. Unlike the usual Save buttons on other Genstat menus, the columns to be saved must be specified before using the Run button. In the Store dialog shown to the right, every column possible has been specified, and the option to display the results in a spreadsheet has been selected. Clicking the **Run** button will perform the analysis and produce the following spreadsheet. Note as no dye-bias has been estimated this column will not appear in the spreadsheet.

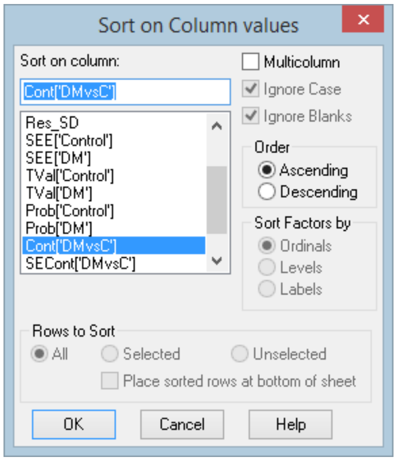

Microarray Estimates from Log-ratios Store O...

Save
- ☑ Probe/Gene IDs In: Probes
- ☑ Estimates In: Est
- ☑ Standard Errors of Estimates In: SEE
- ☑ Residual Degrees of Freedom In: DF
- ☑ Residual Standard Deviation In: Res_SD
- ☑ Dye Bias Effects In: DyeBias
- ☑ t-values In: TVal
- ☑ Probabilities of Estimates In: Prob
- ☑ Contrast Estimates In: Cont
- ☑ Standard Errors of Contrasts In: SECont
- ☑ Contrast t-values In: CTVal
- ☑ Probabilities of Contrast Estimates In: CPVal
- ☑ Display in Spreadsheet


Spreadsheet [Book;4]

Row	Probes	Est['Control']	Est['DM']	DF	Res SD	SEE['Control']	SEE['DM']	TVal['Control']	TVal['DM']	Prob['Control']	Prob['DM']
1	000110ZB001771	0.0401478	-0.0401478	3	0.151327	0.0378318	0.0378318	1.06122	-1.06122	0.366453	0.366453
2	000110ZB001775	-0.13128	0.13128	3	0.236376	0.0590939	0.0590939	-2.22154	2.22154	0.112881	0.112881
3	000110ZB001781	0.0563104	-0.0563104	3	0.314959	0.0787398	0.0787398	0.715145	-0.715145	0.526151	0.526151
4	000110ZB001791	-0.159339	0.159339	3	0.214984	0.0537461	0.0537461	-2.96466	2.96466	0.0593214	0.0593214

This spreadsheet contains a line for each probe, and columns for the different targets and contrasts. The spreadsheet is in probe order but can be sorted into various orders using the Spread | Sort menu so that for example all the most differentially expressed probes are at the start or end of the spreadsheet. The Sort dialog (right) shows the spreadsheet being sorted on the values of the Contrast (DM vs. C). The most differentially under-expressed probes will be at the start of the spreadsheet, and those over-expressed will be at the end. This spreadsheet has been saved as “Estimates13-6-9.gsh”.


Sort on Column values

Sort on column: Cont['DMvsC']

Res_SD
SEE['Control']
SEE['DM']
TVal['Control']
TVal['DM']
Prob['Control']
Prob['DM']
Cont['DMvsC']
SECont['DMvsC']

- □ Multicolumn
- ☑ Ignore Case
- ☑ Ignore Blanks

Order
- ● Ascending
- ○ Descending

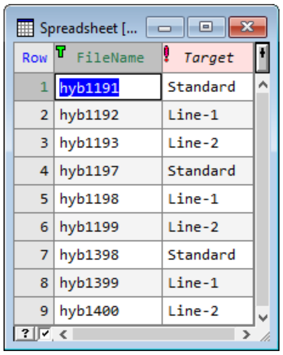
Sort Factors by
- ● Ordinals
- ○ Levels
- ○ Labels

Rows to Sort
- ● All ○ Selected ○ Unselected
- □ Place sorted rows at bottom of sheet

One channel ANOVA

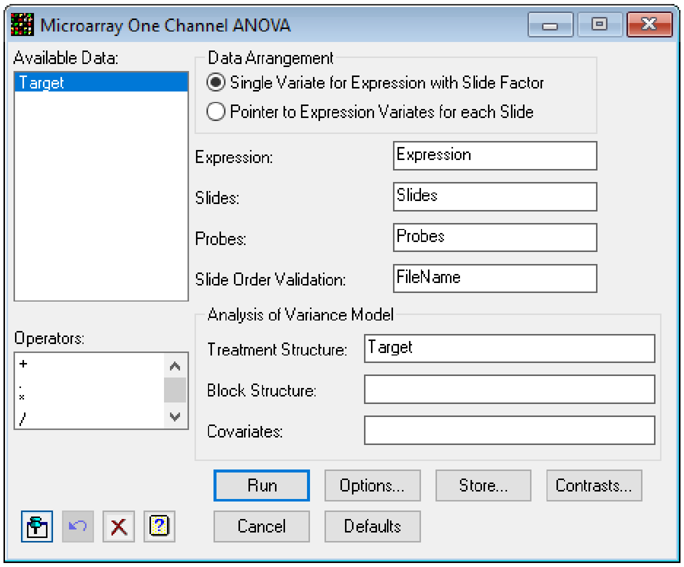
The One Channel ANOVA menu performs an analysis of variance on all the probes/genes in parallel. It assumes a single value from each slide, as with Affymetrix chips. The treatment structure across the slides needs to be provided in a small spreadsheet that supplements the main spreadsheet. A wide range of statistics from the ANOVA can be saved on each gene, and these can be saved in a spreadsheet.

Using the expression values from “Hyb-Expressions.gsh”, we need a spreadsheet that gives the treatments on the various slides. This has been entered into the spreadsheet to the right, which only contains two columns, FileNames – the labels of the slides in the main spreadsheet, and Target – the targets (treatments) applied to the 9 slides. This experiment had three replicates of three treatments, a standard control line and two experimental lines that were to be compared with the control. This Slide information spreadsheet has been saved as “HybFiles.gsh”.



Row	FileName	Target
1	hyb1191	Standard
2	hyb1192	Line-1
3	hyb1193	Line-2
4	hyb1197	Standard
5	hyb1198	Line-1
6	hyb1199	Line-2
7	hyb1398	Standard
8	hyb1399	Line-1
9	hyb1400	Line-2

Opening this menu using Stats | Microarrays | Analyse | One Channel ANOVA menu, we would complete this as to the right to analyse this experiment. The columns from the main expression file are entered into the Expression, Slides and Probes fields. The column FileName is entered into the Slide Validation field and is used to check that the treatment factors correspond to the order of the slides in the main expression spreadsheet. If the labels do not match between the Slides factor and the Slide Order Validation entries, this will be flagged as an error in the output. If the order between the labels and the Slide Order Validation entries do not match, then a warning will be printed in the output and the Treatment factors used in the Treatment and Block Structure fields will be sorted into the order that matches with the Slides factor. The Slide spreadsheet can contain multiple Treatment and Block factors and the same types of treatment structures can be entered in the Treatment and Block Structure fields as in the standard ANOVA menu. For example, if the slides had been grouped into three replicates, we could have a factor Rep in the Block Structure field to remove between replicate differences. As this trial is being treated as a completely randomized experiment, the Block Structure field is left blank. If we had a factorial arrangement of two treatments (A and B) we could enter A*B as the Treatment Structure, and if we had 4 separate cell lines, two treated with one chemical, and two with another, we could use the nested structure Chemical/CellLine. See the Stats | Analysis of Variance | General menu for more details on treatment and block structures.



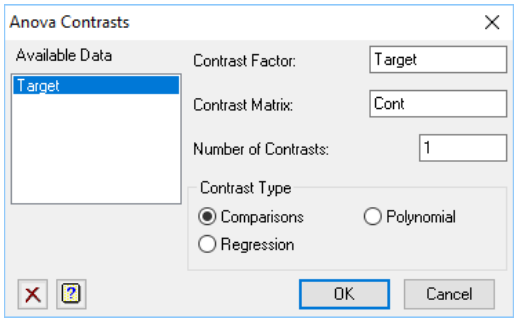
Available Data: Target

Data Arrangement
● Single Variate for Expression with Slide Factor
○ Pointer to Expression Variates for each Slide

Expression: Expression
Slides: Slides
Probes: Probes
Slide Order Validation: FileName

Analysis of Variance Model
Treatment Structure: Target
Block Structure:
Covariates:

Contrasts been the treatment levels can also be defined using the Contrasts button. These are added to the Treatment Structure field as COMP, REG, or POL functions. The COMP and REG functions require contrast matrices that are produced when the **OK** button is pressed on the dialog opposite.

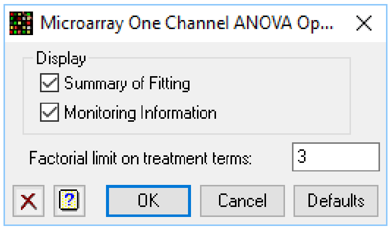


Available Data: Target

Contrast Factor: Target
Contrast Matrix: Cont
Number of Contrasts: 1

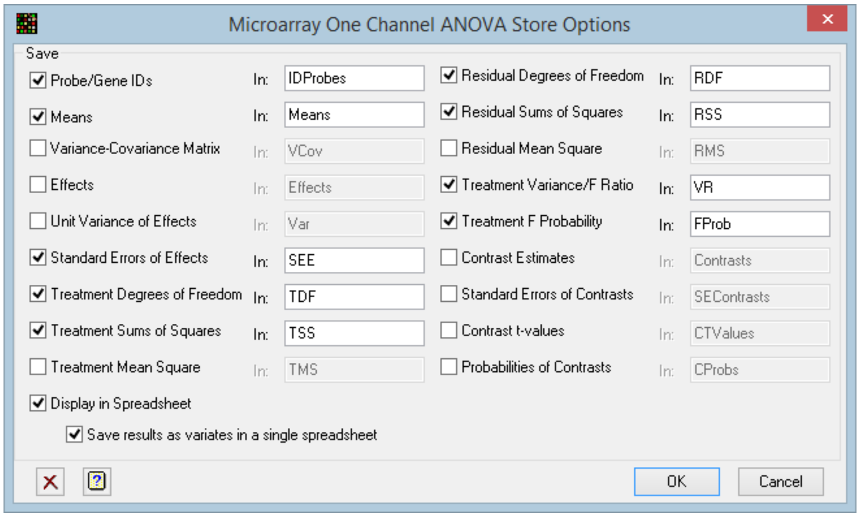
Contrast Type
● Comparisons ○ Polynomial
○ Regression

There are not many options to be set on the dialog opened with the Options button. The options control what is printed in the output window and allow restrictions on order of the factorial terms fitted in the ANOVA. For example, if the factorial limit is set to 2, all third order interactions (e.g., A.B.C) and higher will be omitted from the analysis.

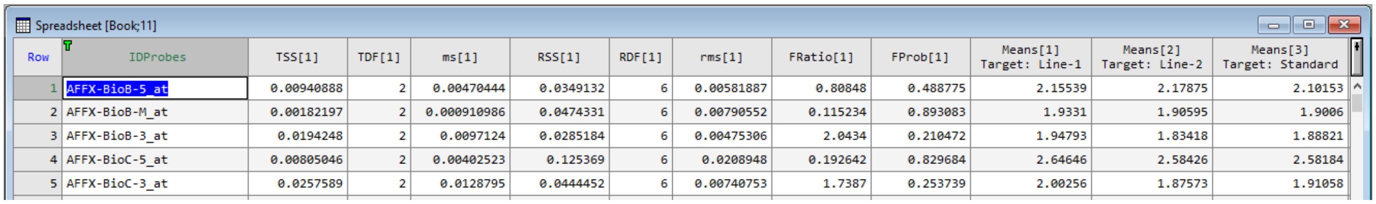


☑ Summary of Fitting
☑ Monitoring Information
Factorial limit on treatment terms: 3

To store the results of the analysis of variances for each probe, use the Store button (this must be set before running the analysis), select the results to be saved, and give names to contain these. The structures that contain results for each treatment level (such as means or effects) will be stored in matrices. The option to display these results as a spreadsheet can be selected, in which case the results will be saved in multiple spreadsheets, unless the supplementary option “Save Results as variates in a Single spreadsheet” is selected. In this case the matrices will be converted to multiple variates, all displayed in one spreadsheet. When the Run button is pressed on the main menu, the following spreadsheet will be produced.



☑ Probe/Gene IDs In: IDProbes
☑ Means In: Means
☐ Variance-Covariance Matrix In: VCov
☐ Effects In: Effects
☐ Unit Variance of Effects In: Var
☑ Standard Errors of Effects In: SEE
☑ Treatment Degrees of Freedom In: TDF
☑ Treatment Sums of Squares In: TSS
☐ Treatment Mean Square In: TMS
☑ Residual Degrees of Freedom In: RDF
☑ Residual Sums of Squares In: RSS
☐ Residual Mean Square In: RMS
☑ Treatment Variance/F Ratio In: VR
☑ Treatment F Probability In: FProb
☐ Contrast Estimates In: Contrasts
☐ Standard Errors of Contrasts In: SEContrasts
☐ Contrast t-values In: CTValues
☐ Probabilities of Contrasts In: CProbs
☑ Display in Spreadsheet
☑ Save results as variates in a single spreadsheet

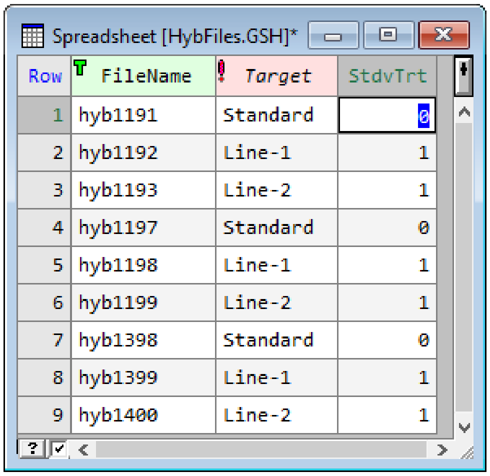

Spreadsheet [Book;11]

Row	IDProbes	TSS[1]	TDF[1]	ms[1]	RSS[1]	RDF[1]	rms[1]	FRatio[1]	FProb[1]	Means[1] Target: Line-1	Means[2] Target: Line-2	Means[3] Target: Standard
1	AFFX-BioB-5_at	0.00940888	2	0.00470444	0.0349132	6	0.00581887	0.80848	0.488775	2.15539	2.17875	2.10153
2	AFFX-BioB-M_at	0.00182197	2	0.000910986	0.0474331	6	0.00790552	0.115234	0.893083	1.9331	1.90595	1.9006
3	AFFX-BioB-3_at	0.0194248	2	0.0097124	0.0285184	6	0.00475306	2.0434	0.210472	1.94793	1.83418	1.88821
4	AFFX-BioC-5_at	0.00805046	2	0.00402523	0.125369	6	0.0208948	0.192642	0.829684	2.64646	2.58426	2.58184
5	AFFX-BioC-3_at	0.0257589	2	0.0128795	0.0444452	6	0.00740753	1.7387	0.253739	2.00256	1.87573	1.91058

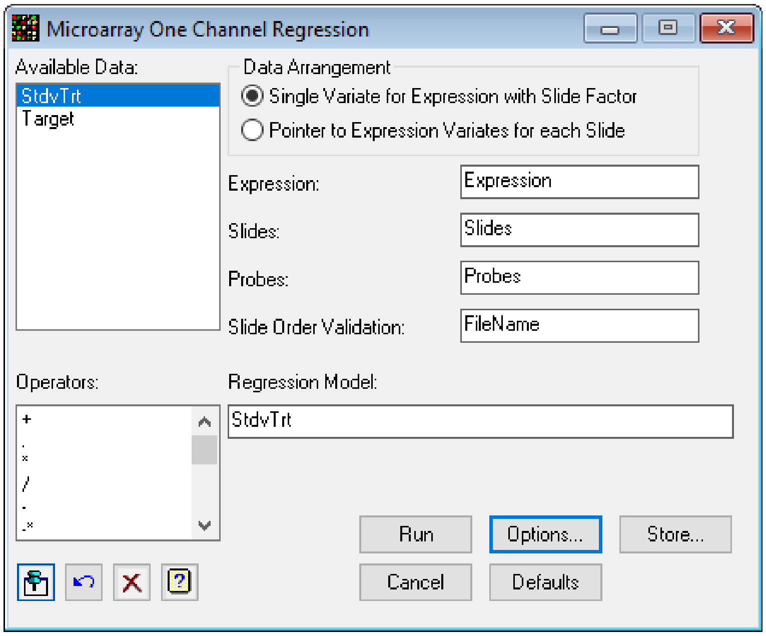
One Channel Regression Analysis

If the model that you wish to fit to the expression data is not balanced, or contains regression terms, then we can use the One Channel Regression Analysis menu. This performs a regression on all the probes/genes in parallel. It assumes a single value from each slide, as with Affymetrix chips. The regression model terms across the slides are provided in a small spreadsheet that supplements the main spreadsheet. A wide range of statistics from the regression can be saved on each gene, and these can be displayed in a spreadsheet.

Opening this menu using Stats | Microarrays | Analyse | One Channel Regression menu, we would complete this as to the right to analyse this experiment for the difference between the control and the average of the two treatments Line-1 and Line-2, by inserting a column called StdvTrt (Spread | Insert Column after current column) and entering the values shown. The columns from the main expression file are entered in the Expression, Slides and Probes fields. The column FileName is entered in the Slide Validation field and is used to check that the regression terms correspond to the order of the slides in the main expression spreadsheet. If the labels do not match between the Slides factor and the Slide Order Validation entries, this will be flagged as an error in the output. If the order between the labels and the Slide Order Validation entries do not match, then a warning will be printed in the output and the treatment terms used in the Regression Model field will be sorted into the order that matches with the Slides factor. The Slide spreadsheet can contain multiple variates and factors to be used in the Regression model. Here a variate with just two values, 0 for control entries and 1 for the treated slides has been put into the model field and fitting this will create an estimate of the difference between the control and the average of Line-1 and Line-2. The **Options** button, which opens the options dialog shown to the below right, can be used to separate out or pool the sums of squares of multiple terms in the regression and control what is printed, and the model terms fitted.



Row	FileName	Target	StdvTrt
1	hyb1191	Standard	0
2	hyb1192	Line-1	1
3	hyb1193	Line-2	1
4	hyb1197	Standard	0
5	hyb1198	Line-1	1
6	hyb1199	Line-2	1
7	hyb1398	Standard	0
8	hyb1399	Line-1	1
9	hyb1400	Line-2	1

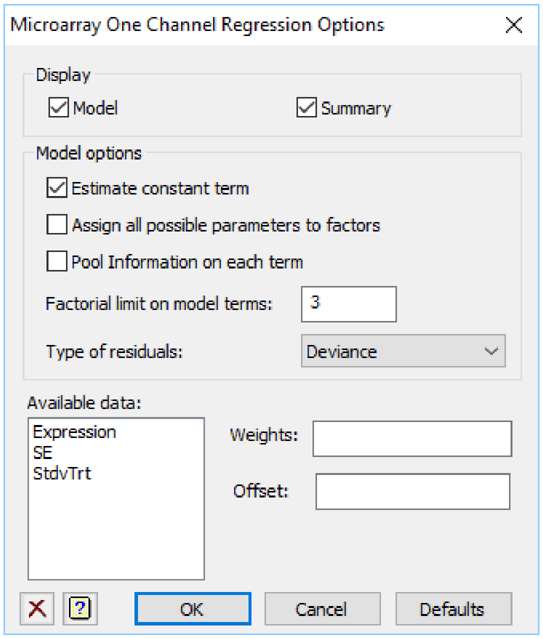


● Single Variate for Expression with Slide Factor
○ Pointer to Expression Variates for each Slide

Expression: Expression
Slides: Slides
Probes: Probes
Slide Order Validation: FileName
Regression Model: StdvTrt

Use the Store button to save the results from the regression and to display these in a spreadsheet. This dialog is shown on the following page. Clicking the **Run** button produces the following output in the Output window and the spreadsheet on the next page.

```
Regression analysis

Response variate:
        Fitted terms:   Constant, StdvTrt

Summary of t and F probabilities

% probes with probabilities in the ranges:

                >10%    5 - 10%    1 - 5%   0.1 - 1%   <= 0.1%

   Constant     0.25      0.23      0.65      1.89      96.99
    StdvTrt    59.70      8.45     13.51     10.40       7.94
    StdvTrt    59.70      8.45     13.51     10.40       7.94
```

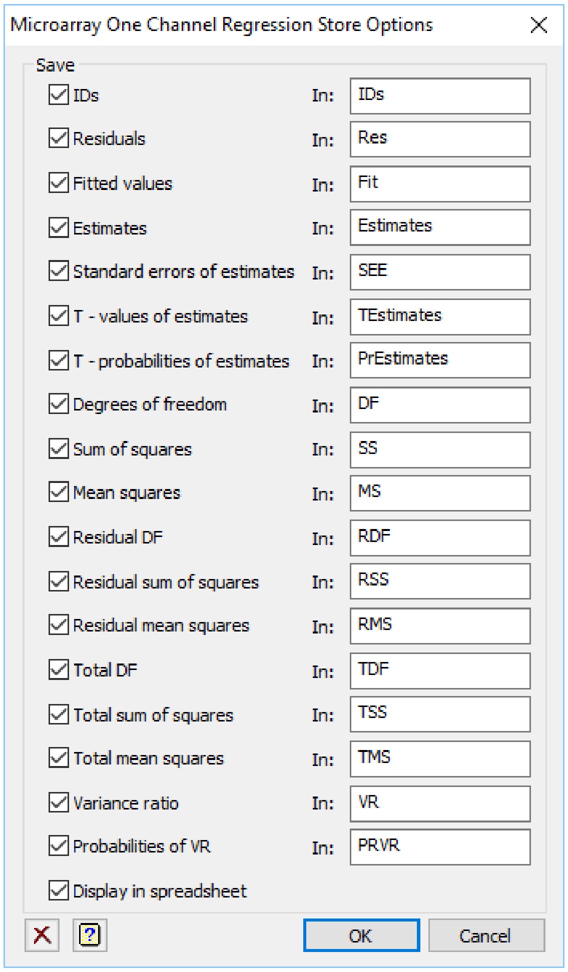


☑ Model ☑ Summary

☑ Estimate constant term
☐ Assign all possible parameters to factors
☐ Pool Information on each term

Factorial limit on model terms: 3
Type of residuals: Deviance
Weights:
Offset:

The spreadsheet will contain multiple pages, as many of the results are saved in matrices and each spreadsheet page can only contain a single matrix. The page titles specify what is contained on each page, and a units' column is added to identify the probes associated with each line.

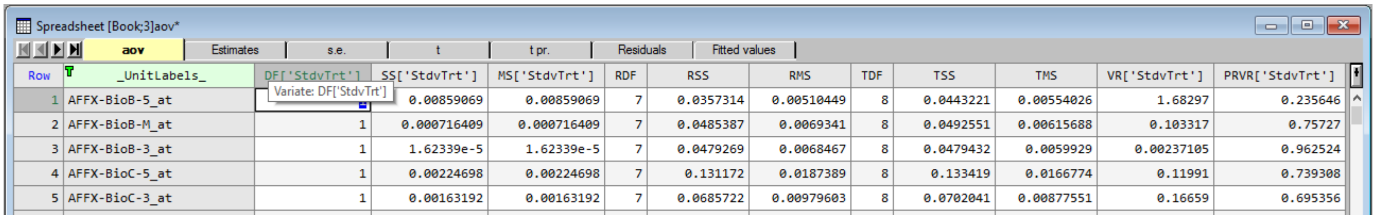


Microarray One Channel Regression Store Options

Save

☑ IDs	In:	IDs
☑ Residuals	In:	Res
☑ Fitted values	In:	Fit
☑ Estimates	In:	Estimates
☑ Standard errors of estimates	In:	SEE
☑ T - values of estimates	In:	TEstimates
☑ T - probabilities of estimates	In:	PrEstimates
☑ Degrees of freedom	In:	DF
☑ Sum of squares	In:	SS
☑ Mean squares	In:	MS
☑ Residual DF	In:	RDF
☑ Residual sum of squares	In:	RSS
☑ Residual mean squares	In:	RMS
☑ Total DF	In:	TDF
☑ Total sum of squares	In:	TSS
☑ Total mean squares	In:	TMS
☑ Variance ratio	In:	VR
☑ Probabilities of VR	In:	PRVR

☑ Display in spreadsheet



Spreadsheet [Book;3]aov*

Row	_UnitLabels_	DF['StdvTrt']	SS['StdvTrt']	MS['StdvTrt']	RDF	RSS	RMS	TDF	TSS	TMS	VR['StdvTrt']	PRVR['StdvTrt']
1	AFFX-BioB-5_at	Variate: DF['StdvTrt']	0.00859069	0.00859069	7	0.0357314	0.00510449	8	0.0443221	0.00554026	1.68297	0.235646
2	AFFX-BioB-M_at	1	0.000716409	0.000716409	7	0.0485387	0.0069341	8	0.0492551	0.00615688	0.103317	0.75727
3	AFFX-BioB-3_at	1	1.62339e-5	1.62339e-5	7	0.0479269	0.0068467	8	0.0479432	0.0059929	0.00237105	0.962524
4	AFFX-BioC-5_at	1	0.00224698	0.00224698	7	0.131172	0.0187389	8	0.133419	0.0166774	0.11991	0.739308
5	AFFX-BioC-3_at	1	0.00163192	0.00163192	7	0.0685722	0.00979603	8	0.0702041	0.00877551	0.16659	0.695356

Robust means analysis

The Robust Means Analysis menu produces medians of the probe effects over the slides for a single treatment factor, using the same algorithm used in RMA. This is an iterative analysis that removes the median slide effects in estimating the median level of each probe over all the slides. Unless you have a single treatment in your experiment (unlikely), this menu will be of little use, but is made available in case you want to use the algorithm used in the full RMA analysis. You would probably get more power out of your experiment using the One Channel ANOVA menu.

Empirical Bayes error estimation

With the large number of genes analysed in parallel on the same series of slides, the variation in the results for each gene may be thought of as coming from a common error distribution. If all the results were generated from a normal error process, we would expect the distribution of standard deviations for each gene to follow a Chi-square distribution. If this was the case, considerable extra power could be obtained if we model the genes together, borrowing information from the whole distribution of standard deviations. The empirical Bayes error estimation does this by modelling the distribution of the standard deviations of the results over all probes. The distribution of standard deviations has two components, a single common standard deviation of the uniform error process operating on all genes, and a specific component of variance unique to each gene. A prior distribution for the standard deviations, or equivalently, the variances, is assumed. In this approach, it is assumed that the reciprocal of the variance is distributed with a multiple of a Chi-square distribution with d_0 degrees of freedom, i.e.

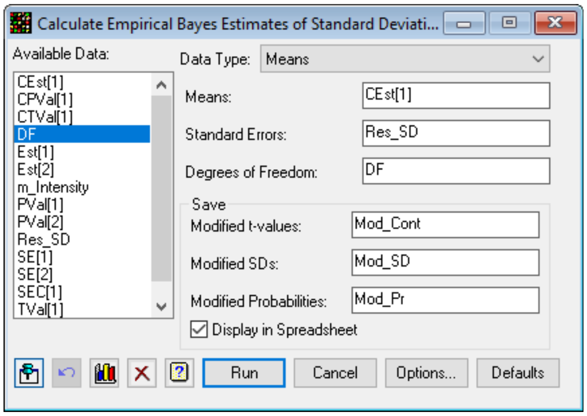
$$\frac{1}{s_p^2} \sim \frac{1}{d_0 s_0^2}\chi^2_{d_0}$$

If the parameters of this distribution, the prior degrees of freedom and standard deviation, $\boldsymbol{d_0}$ and $\boldsymbol{s_0}$ are estimated, more information can be gained on an individual probe, by shrinking it towards the prior standard deviation, $\boldsymbol{s_0}$. The relative amount of information in the prior and individual standard deviation of a probe, ($\boldsymbol{s_0}$ and $\boldsymbol{s_p}$ respectively) is specified by their degrees of freedom, $\boldsymbol{d_0}$ and $\boldsymbol{d_p}$. The modified standard deviation, $\tilde{s}_p$, is then given by the weighted average of $\boldsymbol{s_0}$ and $\boldsymbol{s_p}$:

$$\tilde{s}_p = \sqrt{\frac{d_0 s_0^2 + d_0 s_p^2}{d_0 + d_p}}$$

A modified *t*-test can then be performed using the modified standard deviation with d_0+d_p degrees of freedom. The method can also produce the p-values from a test of the differential expression being different from zero.

Using the estimates from the 13-6 to 13-9 series (saved in “Estimates13-6-9.gsh”), we can create modified *t*-statistics and p-values for the contrast effects of DM vs. Control. Opening the menu Stats | Microarrays | Analyse | Empirical Bayes Error Estimation gives us the window to the right when the fields are filled in with the appropriate column names. The Data Type dropdown list allows the data to be given in 3 formats, means (as in this example), T values, or a Pointer to a set of columns from which means and standard deviations are calculated from each row over the set of columns. Note: a pointer is a Genstat structure that specifies a list of data structures to be treated as a group, that can be defined using the View | Data View menu). The resulting columns are specified in the Save section, with the option of adding these back to the source spreadsheet if it is still open.


Calculate Empirical Bayes Estimates of Standard Deviati...

Available Data:
CEst[1]
CPVal[1]
CTVal[1]
DF
Est[1]
Est[2]
m_Intensity
PVal[1]
PVal[2]
Res_SD
SE[1]
SE[2]
SEC[1]
TVal[1]

Data Type: Means

Means: CEst[1]

Standard Errors: Res_SD

Degrees of Freedom: DF

Save

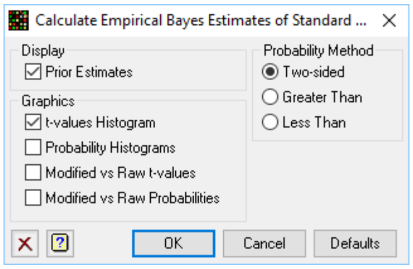
Modified t-values: Mod_Cont

Modified SDs: Mod_SD

Modified Probabilities: Mod_Pr

☑ Display in Spreadsheet

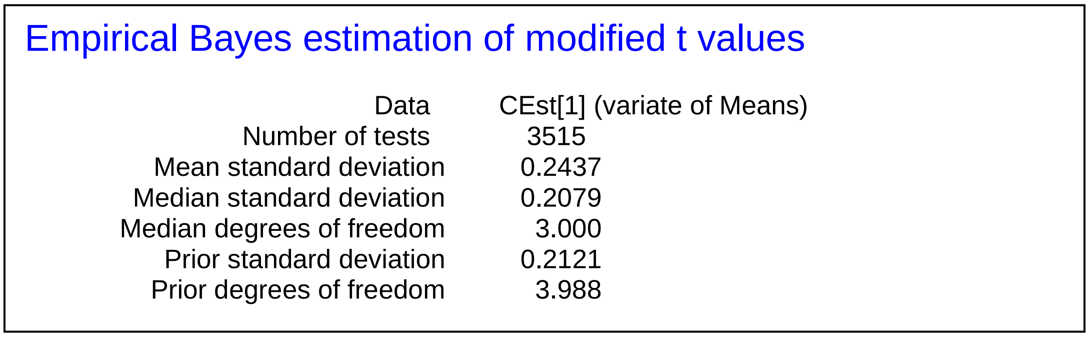
Clicking the **Options** button opens the dialog to the right, which allows you to specify whether the output printed in the Output window, which graphs are plotted, and the nature of the t-test performed (two sided or either of the one sided tests). Here, a two-sided test is used with output of the results to the Output window, and just the histogram of the *t*-values before and after adjustment using the estimated prior parameters. Clicking the **Run** button creates the following output and graphs:


Calculate Empirical Bayes Estimates of Standard ...

Display
☑ Prior Estimates

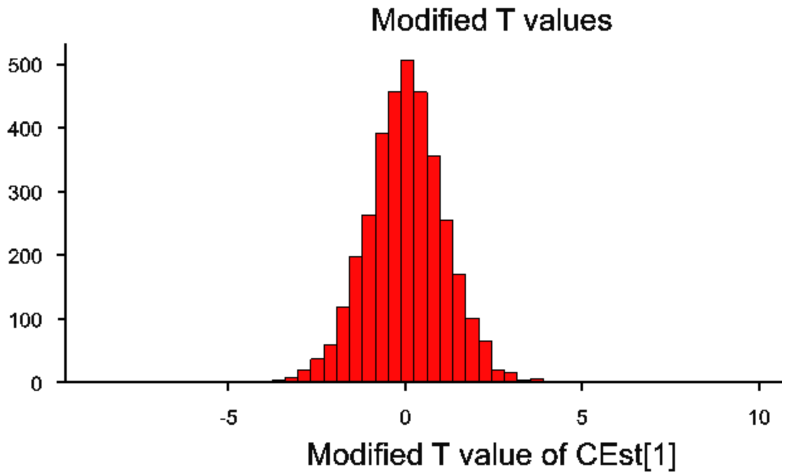
Graphics
☑ t-values Histogram
□ Probability Histograms
□ Modified vs Raw t-values
□ Modified vs Raw Probabilities

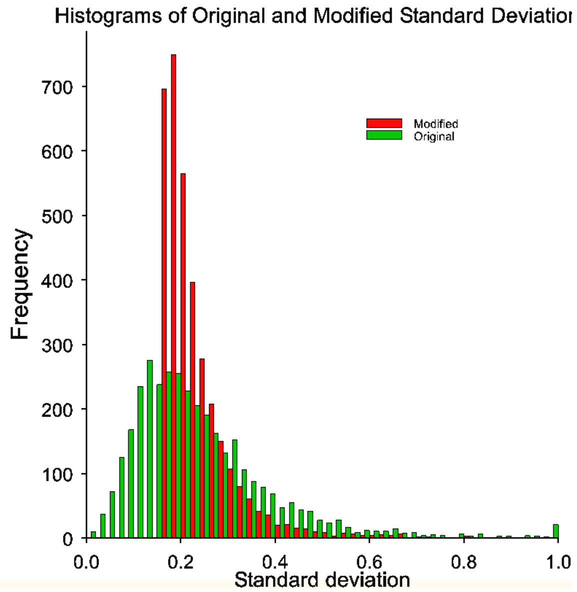
Probability Method
● Two-sided
○ Greater Than
○ Less Than

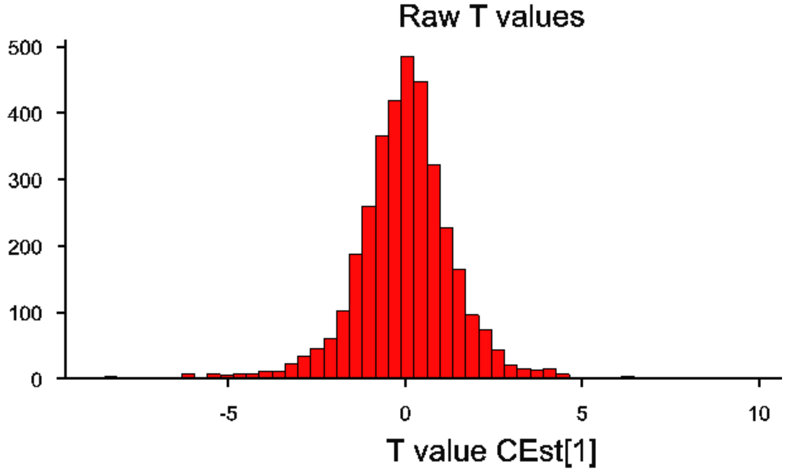


```
Empirical Bayes estimation of modified t values

                       Data       CEst[1] (variate of Means)
            Number of tests         3515
    Mean standard deviation       0.2437
  Median standard deviation       0.2079
 Median degrees of freedom         3.000
   Prior standard deviation       0.2121
  Prior degrees of freedom         3.988
```

The histogram of P values (right) shows overall that the *t*-values have been shrunken towards zero, with the extreme outliers in the *t*-values seeming to be caused by those probes that have very small standard errors. The adjustment towards a common standard deviation has increased the degrees of freedom for the *t*-statistic from 3 to approximately 7. If we plot histograms of the raw and adjusted standard deviations together (below left), we can see how the modified standard deviations (in red) have been shrunk towards the overall mean.







False discovery rate

With so many significance tests on all the probes being performed together, many the tests will be significant by chance, so that we would expect 5% of the tests to be significant at the 5% level even if there were no differential expression for any probes. A better way of understanding how successful the experiment has been to separate differentially expressed probes from non-differentially expressed probes is to estimate the false discovery rate for different levels of significance. The false discovery rate is an estimate of the proportion of non-differentially expressed probes among the tests that are significant at the given level. If this is low, this means that the experiment has been able to effectively separate the differentially expressed probes from the rest. There are two methods of estimating the FDR, one based on a mixture model, and another based on non-parametric methods.

False Discovery Rate using Bonferroni Method

This menu can be used to estimate false discovery rates (defined in the table below) using a Bonferroni-type procedure. This is a non-parametric approach, where for each value of lambda; the observed proportion of the sample that is not differentially expressed (π_0) is calculated. The procedure uses two methods to get an overall measure of π_0. The first uses bootstrapping to choose the value of π_0 which minimises the mean squared error, and the second uses a spline smoother to smooth the values of π_0 around the maximum value of lambda. Unadjusted q-values are then calculated from the estimate of π_0 as π_0*p*(Proportion of tests < p) (where p is the test probability) for each test value, and then the Bonferroni q values are defined as the minimum of the q values above each test value, stepping this procedure down through the sorted p values.

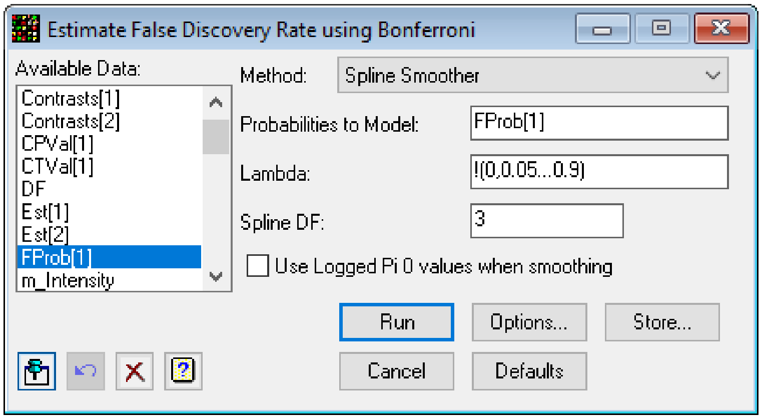
The following table defines some random variables related to m hypothesis tests:

Significance Test	# declared non-significant	# declared significant	Total
# true null hypotheses	U	V	m_0
# non-true null hypotheses	T	S	$m_1 = m - m_0$
Total	$W = m - R$	R	m

m_0 is the number of true null hypotheses.
$m - m_0$ is the number of false null hypotheses.
U is the number of true negatives.
V is the number of false positives.
T is the number of false negatives.
S is the number of true positives.
$H_1 \ldots H_m$ are the null hypotheses being tested.

In m hypothesis tests of which m_0 are true null hypotheses, R is an observable random variable, and S, T, U, and V are unobservable random variables. The proportion of tests that are truly null, π_0, is m_0 divided by m. The false discovery rate (FDR), also known as the q-value of a test, is a commonly used error measure in multiple-hypotheses, defined as FDR = $E(V/R \mid R > 0) \times Pr(R > 0)$, i.e. the expected proportion of false positives findings among all the rejected hypotheses multiplied by the probability of making at least one rejection; the FDR is zero when $R = 0$. Similarly, the false rejection rate (FRR) is defined as FRR = $E(T/W \mid W > 0) \times Pr(W > 0)$, i.e. the expected proportion of false negatives findings among all the accepted hypotheses times the probability of accepting at least one test. We also define the power to be equal to $E(S/m_1 \mid m_1 > 0) \times Pr(m_1 > 0)$.

Opening the Stats | Microarray | Analyse | False Discovery Rate by Bonferroni menu gives the menu to the right. Using the F probability values in the file "Hyb-ANOVA.gwb" in the Empirical Bayes section above, we can fit obtain the esimated false discovery rates to FProb[1].

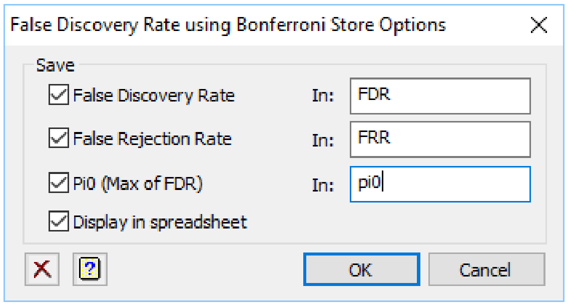


Estimate False Discovery Rate using Bonferroni

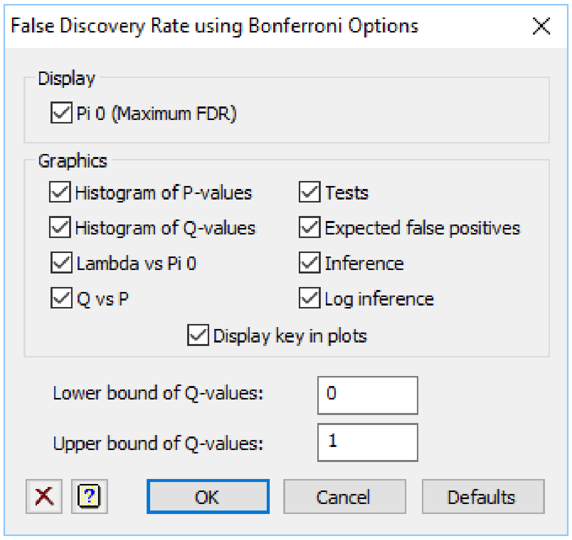
Available Data:
Contrasts[1]
Contrasts[2]
CPVal[1]
CTVal[1]
DF
Est[1]
Est[2]
FProb[1]
m_Intensity

Method: Spline Smoother
Probabilities to Model: FProb[1]
Lambda: !(0,0.05...0.9)
Spline DF: 3
□ Use Logged Pi 0 values when smoothing

The options and results to be saved are set with the Options and Store dialogs shown next.


False Discovery Rate using Bonferroni Store Options

Save
- ☑ False Discovery Rate In: FDR
- ☑ False Rejection Rate In: FRR
- ☑ Pi0 (Max of FDR) In: pi0
- ☑ Display in spreadsheet


False Discovery Rate using Bonferroni Options

Display
- ☑ Pi 0 (Maximum FDR)

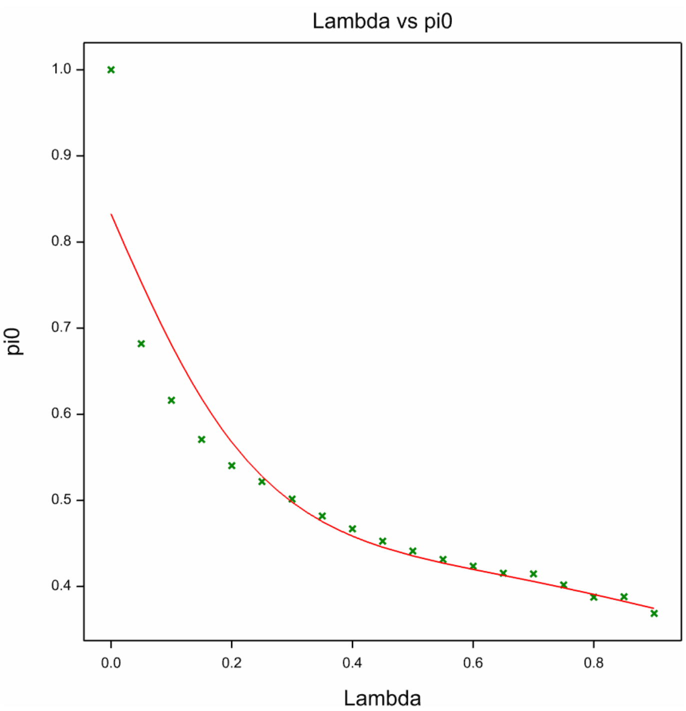
Graphics
- ☑ Histogram of P-values
- ☑ Histogram of Q-values
- ☑ Lambda vs Pi 0
- ☑ Q vs P
- ☑ Tests
- ☑ Expected false positives
- ☑ Inference
- ☑ Log inference
- ☑ Display key in plots

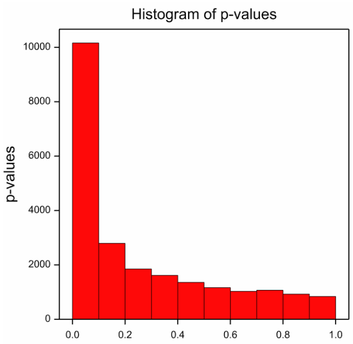
Lower bound of Q-values: 0

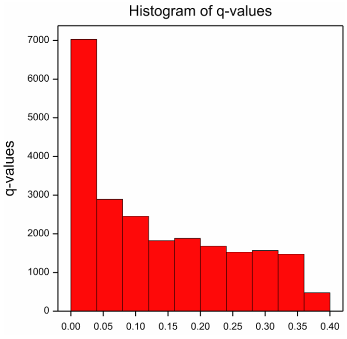
Upper bound of Q-values: 1

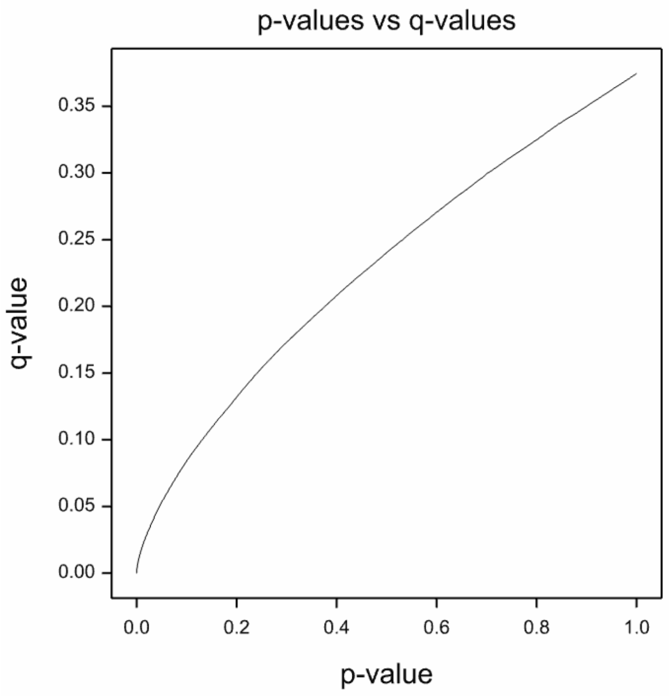
The output contains the estimate of π_0 of 0.3745.

The plot of π_0 vs. lambda shows how the estimate of π_0 flattens off around lambda = 0.4. The estimate of π_0 is taken at the maximum value of lambda, i.e. at lambda = 0.9. An alternative estimate of π_0 could be obtained by reducing the largest value of lambda in the menu.

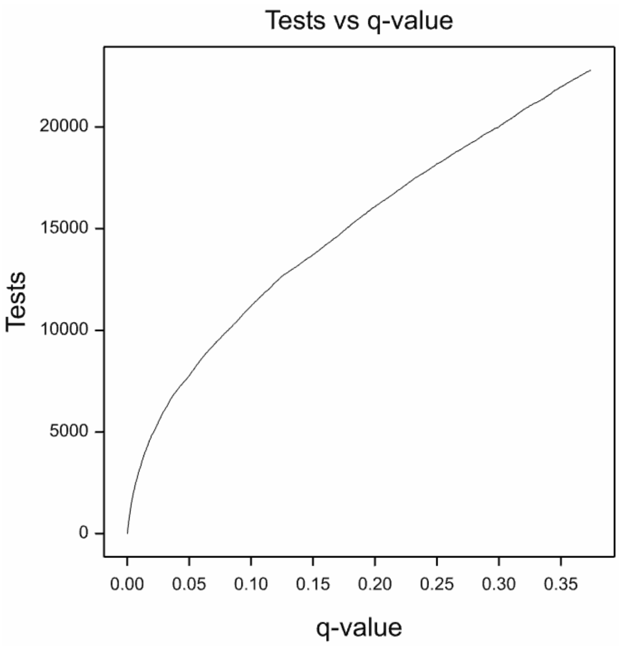


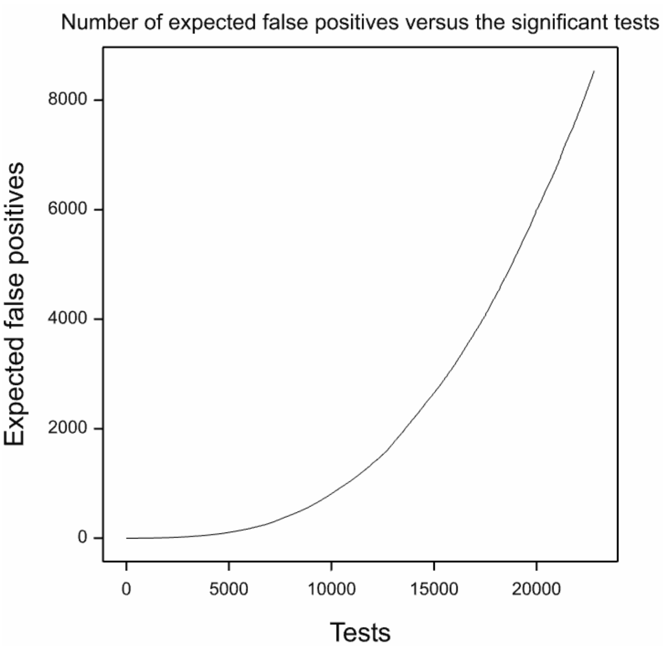


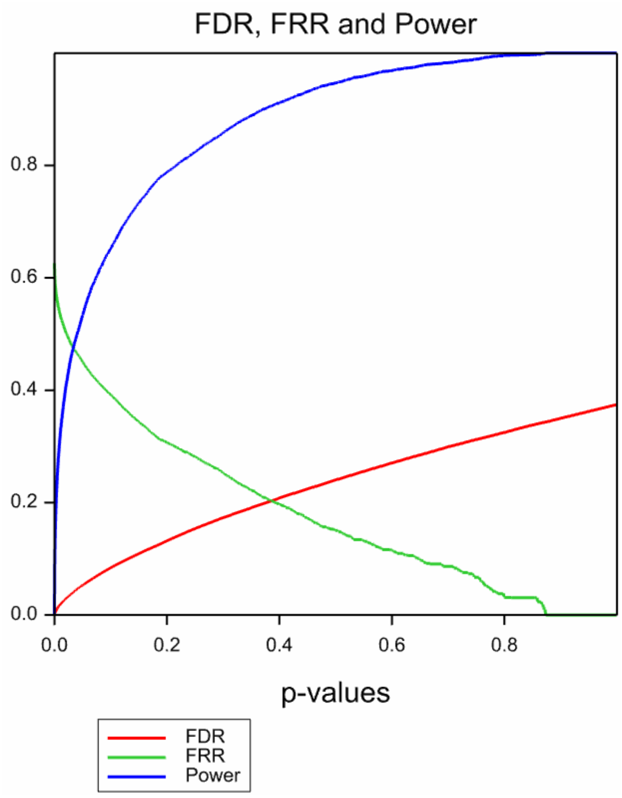


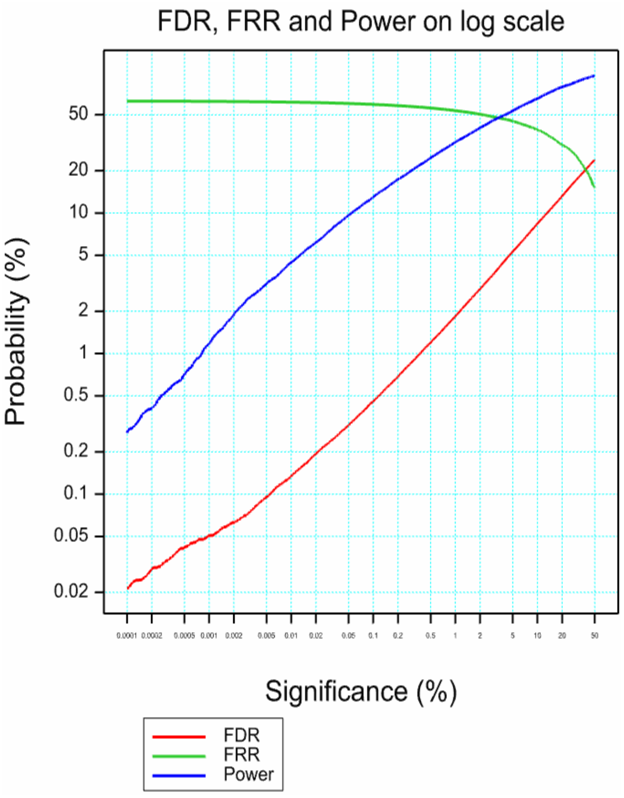


Histograms of the p-values and the resulting q-values are also plotted (above right) which show the reduction in significance when we use q-values. The plot of the q-values vs. the p-values also shows this reduction more clearly. The plots of the number of tests vs. the q-values and the number of expected false positives are also provided (see next page), and these could be used to select a q-value to give an expected number of false positives. Finally, the false discovery and rejection rates are plotted (see next page) on both the normal and log scale so that small probabilities can be examined in more detail. These can be compared with the same graphs produced by the mixture model for estimating false discovery rates in the next section.

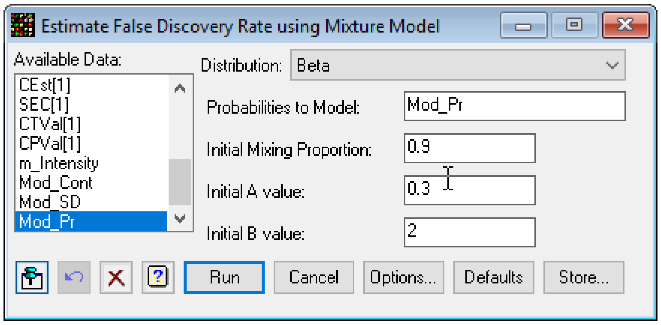

Tests vs q-value
Tests
20000
15000
10000
5000
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
q-value


Number of expected false positives versus the significant tests
Expected false positives
8000
6000
4000
2000
0
0
5000
10000
15000
20000
Tests


FDR, FRR and Power
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
p-values
FDR
FRR
Power


FDR, FRR and Power on log scale
Probability (%)
50
20
10
5
2
1
0.5
0.2
0.1
0.05
0.02
0.0001
0.0002
0.0005
0.001
0.002
0.005
0.01
0.02
0.05
0.1
0.2
0.5
1
2
5
10
20
50
Significance (%)
FDR
FRR
Power

False Discovery Rate using Mixture Model



Distribution: Beta

Probabilities to Model: Mod_Pr

Initial Mixing Proportion: 0.9

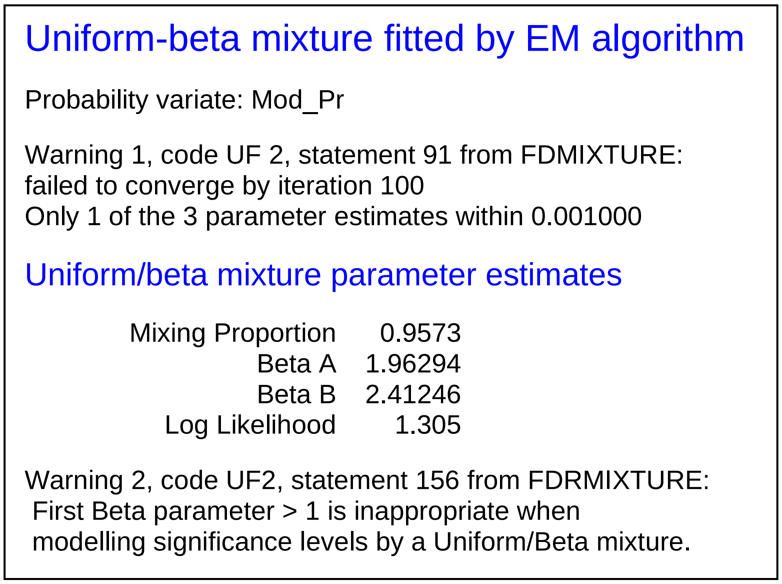
Initial A value: 0.3

Initial B value: 2

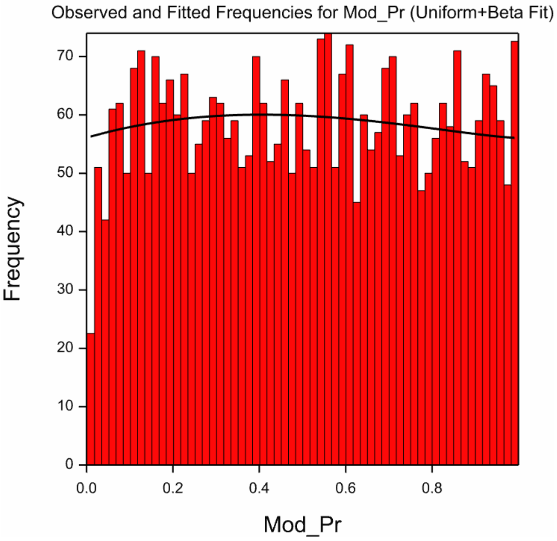
The False Discovery Rate menu can be used to fit a mixture model to a distribution of probabilities. The two components of the mixture can be thought of as those probes which are showing differential expression (modelled by the Beta/Gamma component with probabilities shifted towards zero) and those not responding (the Uniform component, whose probabilities values then form a random sample from the null/uniform distribution of the test statistic). The context is multiple testing, with data from any situation (microarrays here, but also metabolomics and proteomics, among others) where the same simple null-hypothesis, H_0, is tested many times. These tests generate many significance values, which under H_0 have a Uniform distribution, and under the alternative hypothesis, H_a, can be modelled as a Beta density. The false discovery rate (FDR), false rejection rate (FRR) and power of the tests (Allison *et al.*, 2002) with a given level of significance can then be estimated from the parameters of two components. The mixture model parameterization takes a proportion P from the Uniform distribution, and (1 - P) from either a Beta or a Gamma distribution.

Opening the Stats | Microarray | Analyse | False Discovery Rate by Mixture menu gives the menu to the right. Using the Modified p-values generated in the Empirical Bayes section above, we can fit the mixture model, providing some initial parameter estimates.

The output from this model appears in the Output window as:



```
Uniform-beta mixture fitted by EM algorithm

Probability variate: Mod_Pr

Warning 1, code UF 2, statement 91 from FDMIXTURE:
failed to converge by iteration 100
Only 1 of the 3 parameter estimates within 0.001000

Uniform/beta mixture parameter estimates

            Mixing Proportion     0.9573
                       Beta A    1.96294
                       Beta B    2.41246
               Log Likelihood      1.305

Warning 2, code UF2, statement 156 from FDRMIXTURE:
 First Beta parameter > 1 is inappropriate when
 modelling significance levels by a Uniform/Beta mixture.
```



The are two warnings in this output:

1. The parameter estimation did not converge. This is probably because either the initial values are too far away from the optimal parameters, or because the model does not fit the probabilty distribution well.

2. The beta distribution does not have its mode at zero (A > 1), so that it is an inappropriate distribution for describing a False Discovery rate.

When one examines the probability histogram with the fitted mixture plotted over this (previous page), one can see that in fact it looks as if there are virtually no genes significantly responding. The proportion of genes close to zero are less than one would expect from a purely random set of results, and the small beta component (approximately 5%) is being used to describe the slight bulge around the mid range of probablilities. Thus, it is not worth fitting a False Discovery rate to this set of results, as there are fewer than expected responding probes/genes.

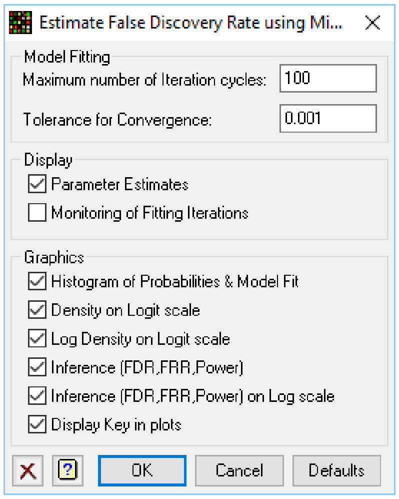
If we return to the Single Channel ANOVA results from the nine Affymetrix chips, saved in the file “Hyb-ANOVA.gwb”, and fit a False Discovery Rate to the F-ratio probabilities in column FProb[1] we get the following output:

```
Uniform-beta mixture fitted by EM algorithm

Probability variate: FProb[1]

Uniform/beta mixture parameter estimates

          Mixing Proportion      0.4166
                     Beta A    0.37154
                     Beta B    3.58105
             Log Likelihood 17005.187
```

In this we see that the estimation has converged, and that a good value of A and B have been fitted. The larger the value of B and the smaller the value of A, the more the p-values are pushed to towards zero. The mixture model only estimates that 41% of the genes are showing no response between the treatment lines 1 & 2 and the control line “Standard”.

The Store button on the menu allows the various estimated results to be saved as usual.

If the estimation does not converge in the default number of iterations you can try changing these using the Options button which opens the dialog to the right. This also controls the graphs plotted and the output. A range of graphs can be plotted after the mixture model has been fitted. These show the fitted model (in three formats to allow close inspection of the quality of the fit) and the estimated FDR, FRR and Power curves. The following graphs come from the fit of the model to FProb[1], as selected to the right.



Model Fitting

Maximum number of Iteration cycles: 100

Tolerance for Convergence: 0.001

Display

☑ Parameter Estimates

☐ Monitoring of Fitting Iterations

Graphics

☑ Histogram of Probabilities & Model Fit

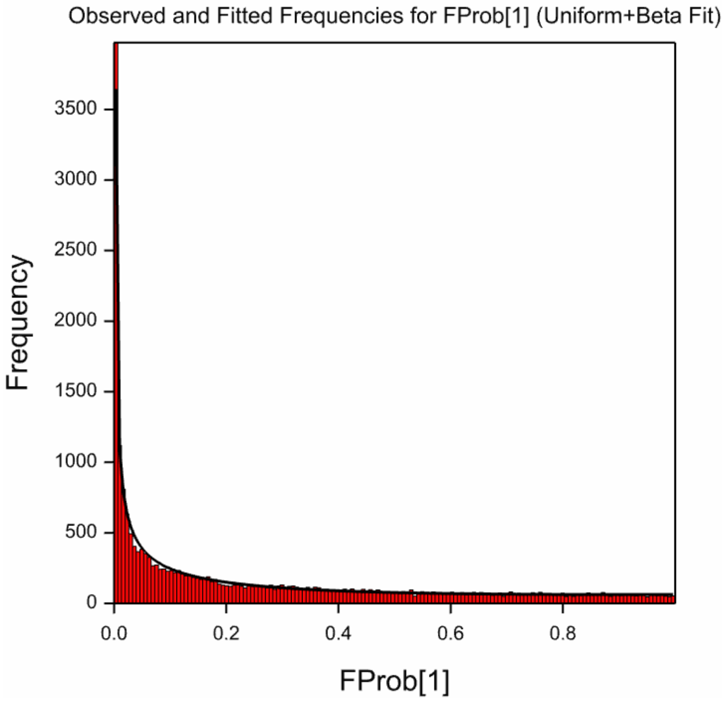
☑ Density on Logit scale

☑ Log Density on Logit scale

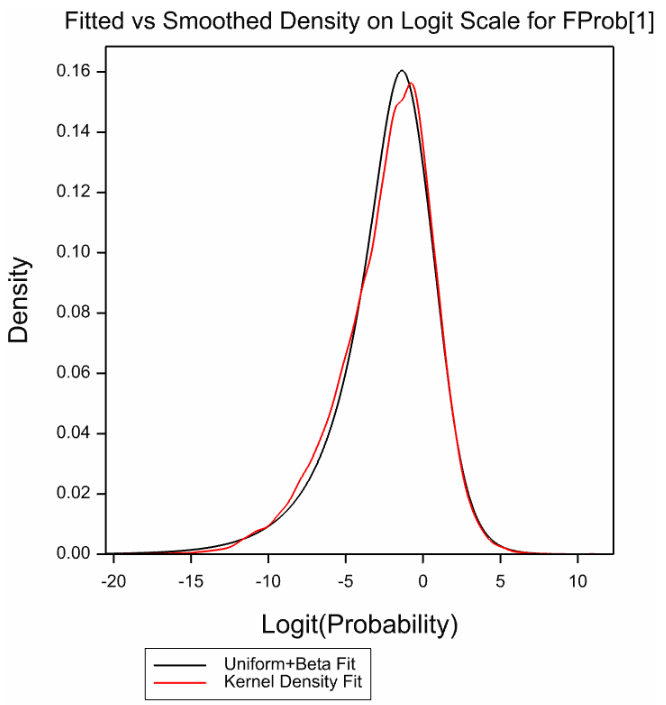
☑ Inference (FDR,FRR,Power)

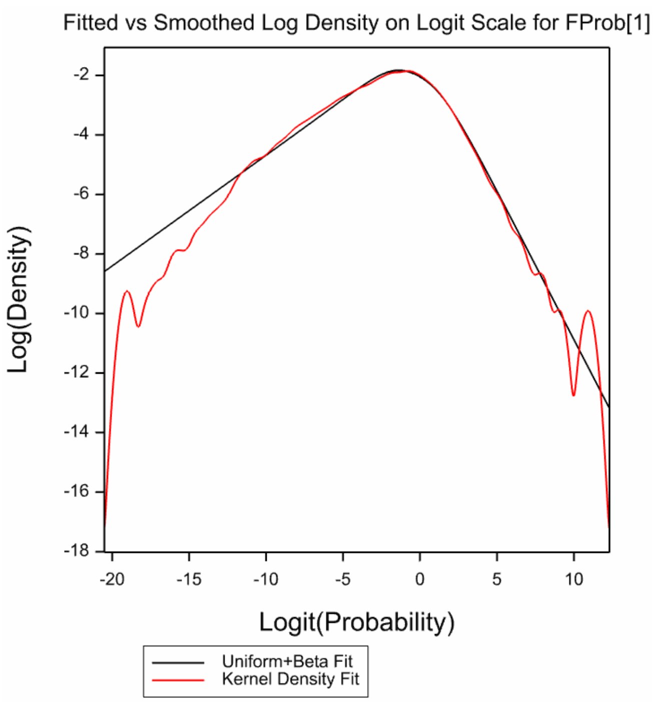
☑ Inference (FDR,FRR,Power) on Log scale

☑ Display Key in plots

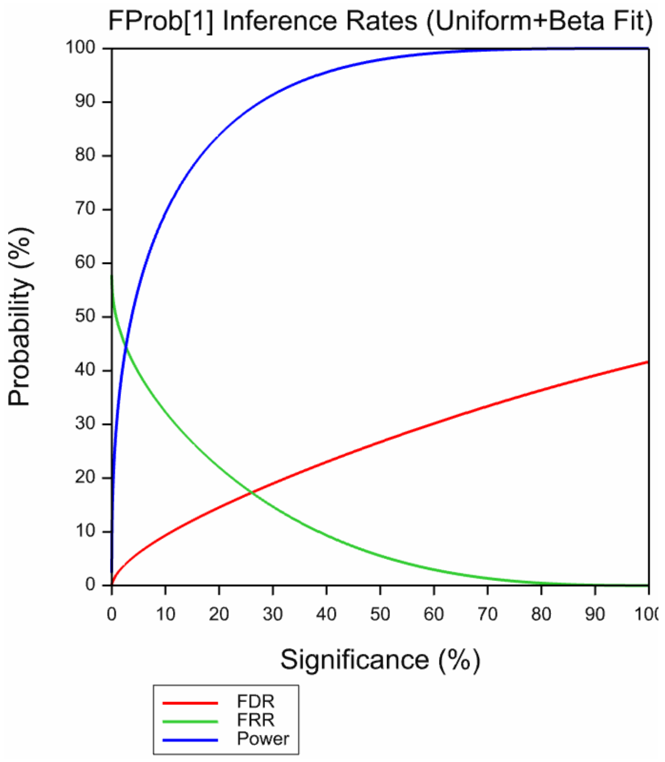


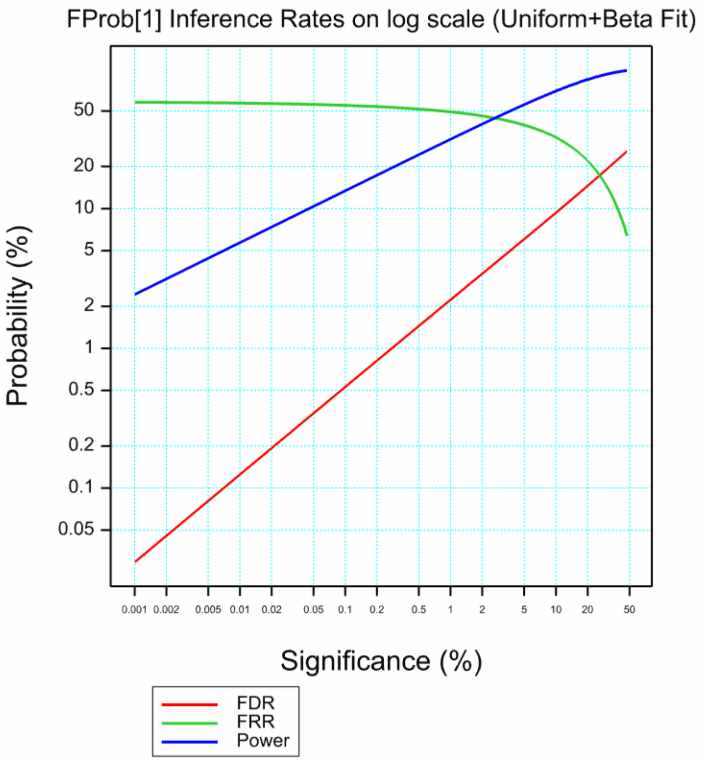
This graph shows most probes/genes have very small p-values. As so many of the values lie close to zero, it is hard to examine the goodness of fit of the model to the data over values less than 0.5. The second graph (below left) produced by the FDR menu (the Density on Logit scale option) rescales the x-axis onto a logit scale, i.e. plotting log(p/1-p) rather than p. This expands the scale near zero and one. The third graph rescales the y-axis using a log scale for the density in the logit plot to examine in even more detail what is happening in the tails of the distribution (the Log Density on Logit scale option). This gives the graphs on the next page.





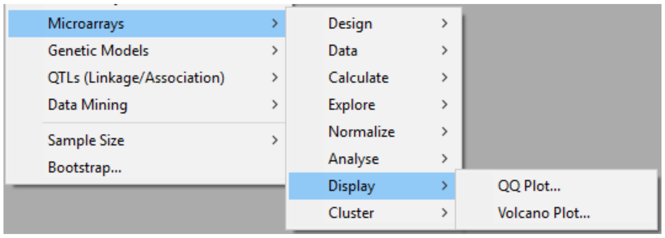
The final two graphs below plot the FDR, FRR and Power curves versus the level of significant used in choosing the genes said to be differentially expressed. In these plots, it can be seen that this experiment has high power and reasonably low False Discovery rates. As there are so many genes classified as responding (~60%), the False Rejection rates are high for small significance values. The second FDR graph plots these curves on a log scale so that values close to zero can be read more accurately.





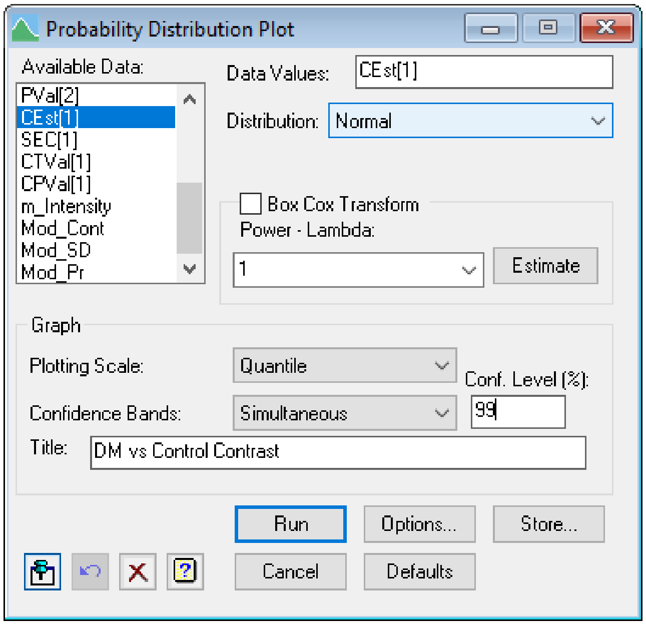
In addition, an alternative False Discovery Rate menu, is available, which does not use a distribution model for the p-values but estimates the False Discovery rates using a Bonferroni type procedure as detailed in Storey (2002). This will give the same type of graphs as the above, plus some others showing how the estimated proportion of non-differentially express genes/probes is estimated.

Display microarray results



The Display menu contains two menu items that allow you to display the results over all the genes. The QQ Plot menu displays the results plotted against their expected scores from a specified distribution (usually a normal or t distribution). The volcano plot displays a scatter plot of a measure of differential expression against a measure of significance, with the option of colouring the points on a third measure.

Display QQ plot



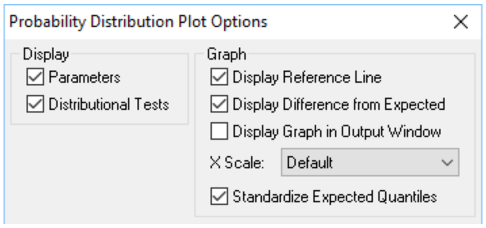
Data Values: CEst[1]
Distribution: Normal
□ Box Cox Transform
Power - Lambda: 1
Plotting Scale: Quantile
Confidence Bands: Simultaneous
Conf. Level (%): 99
Title: DM vs Control Contrast

This menu opens the same menu as the Graphics | Probability Plot menu, which allows a set of values to be plotted against the expected values from a specified distribution. For example, under the assumption that the results over the probes simply come from a random noise process, the central limit theorem would suggest the probe means would follow a normal distribution. Another distribution that could be used is a *t*-distribution for the *t*-values of specific contrasts, with the appropriate degrees of freedom specified. Note that not all the *t*-values will have the same degrees of freedom, but the graph should give a reasonable approximation if the degrees of freedom are not too variable (i.e., if most probes have no missing values).

The menu opens the window to the right. Specify the column to be plotted in the Data Values field, and select the distribution from the dropdown list, specifying the degrees of freedom if required. The menu here is completed using the column CEst[1], the contrast between DM and Control from the file “Estimates13-6-9.gsh”.

The resulting graph from clicking the **Run** button is displayed on the next page.

In this graph, you can see a departure from the normal distribution in the tails, indicating more extreme values than expected from a normal distribution. The green line is the 1-1 line, giving expected values if the distribution was truly normal, and the red curves are the 95 percent confidence curves around this under the null hypothesis that we are sampling from a normal distribution. The points cross the confidence curves in the tails, particularly in the lower tail. To get a closer look at the departures in the tails, we could use the Options button and select the display **Difference from Expected** option (as shown below) to get the following graph below right.



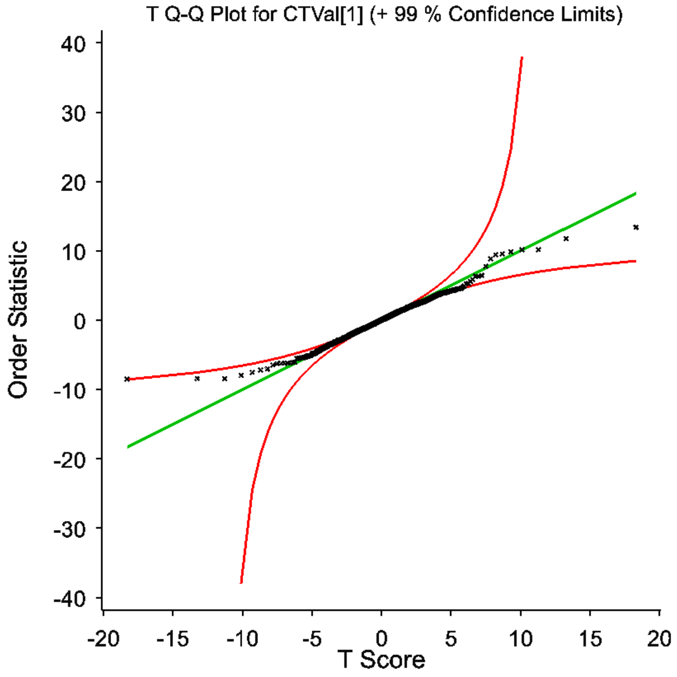
Probability Distribution Plot Options

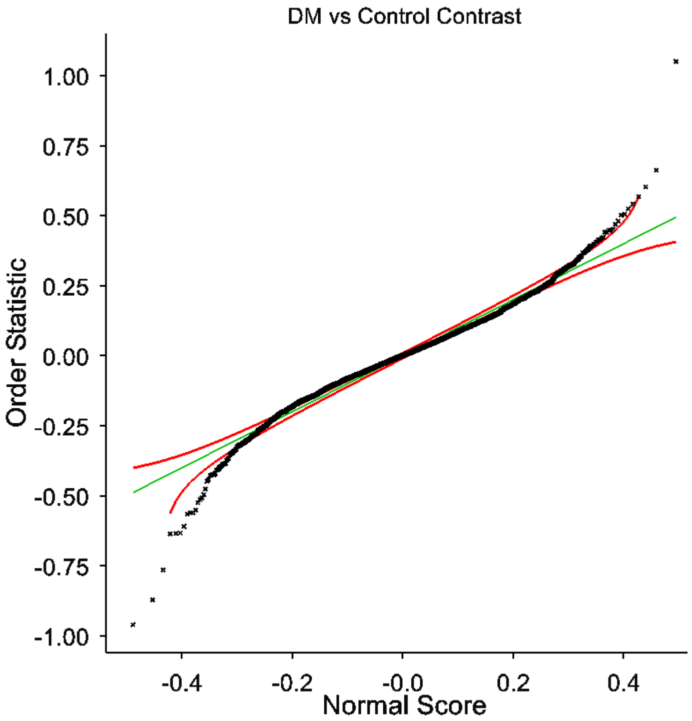
Display
☑ Parameters
☑ Distributional Tests

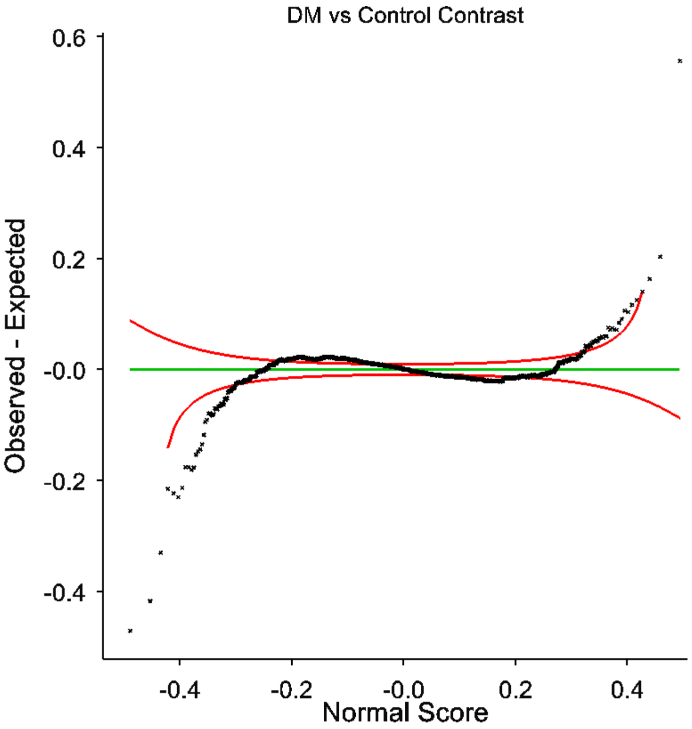
Graph
☑ Display Reference Line
☑ Display Difference from Expected
□ Display Graph in Output Window
X Scale: Default
☑ Standardize Expected Quantiles

This graph allows the more accurate location of where points cross the confidence limits.

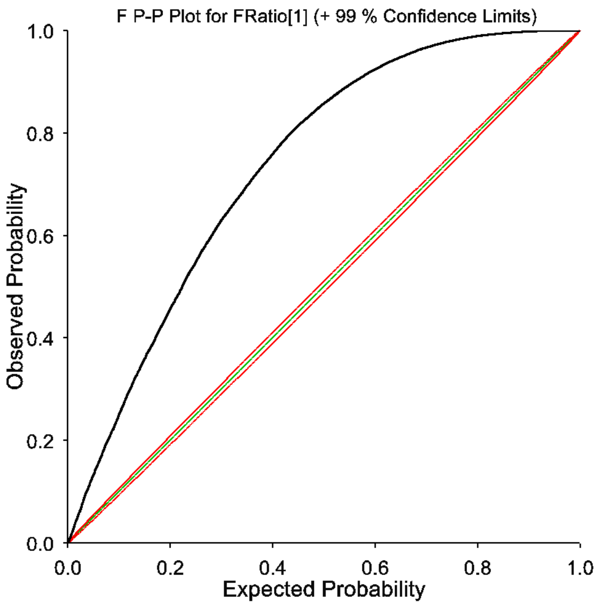
If we now plot the *t*-values in CTVal[1] against the expected t-distribution with 3 degrees of freedom we get the graph below. Note the very wide confidence curves, as a *t*-value with 3 degrees of freedom is quite unstable. Also, note that the points do not cross the confidence limits.





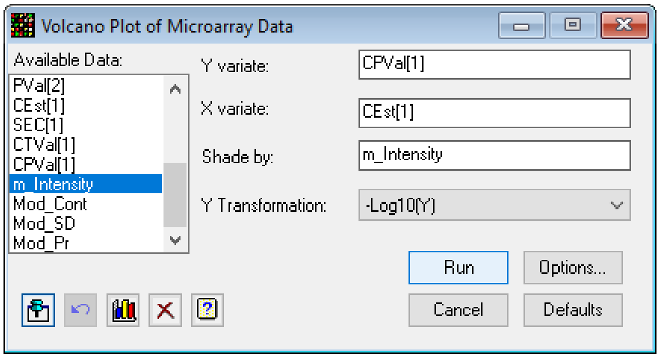


The final example in this section examines the F values in the column FRatio[1] from the file “Hyb-ANOVA.gwb”, which contains the treatment F-ratio for the results from the Single Channel ANOVA for the 9 Affymetrix slides considered earlier. These are compared with the expected values from an F distribution with 2 and 6 degrees of freedom. As F-ratios are very skewed, this comparison is best plotted on a probability scale. The resulting graph is shown to the right, and this shows the same result as found in the False Discovery Rate menu for the same data, i.e. many more of the probes shown significant differential expression than expected under the null hypothesis of no differential expression.



Display volcano plot

This menu shows jointly the level of differential expression and the significance on a single graph. It is called a volcano plot as the points typically take up a v shape looking like the ash spewing out of a volcano. This is because it is rare to get points with very small levels of differential expression, but a high level of significance, and vice versa. In addition, as the significance is only positive, the negative and positive expressions are plotted against positive values, generally giving a symmetric plot (unless there is a big difference in the number of positively and negatively expressed log-ratios). Opening the menu gives the window to the right. In this case, we will look at the CEst[1], the contrast between DM and Control from the file “Estimates13-6-9.gsh”, and CPVal[1], the corresponding probability under a two-sided null hypothesis of no differential expression. To demonstrate the ability to colour the points on another variate, we will use the mean intensity of the spots for a probe over the slides. To add this variate to the Estimates spreadsheet, open the full data set “Data13-6-9.gwb” and use the Spread | Calculate | Summary Stats menu to obtain the means of intensity for each probe. The completed dialog that does this is shown to the right. Clicking **OK** will create a spreadsheet containing the means (shown on the next page).


Volcano Plot of Microarray Data

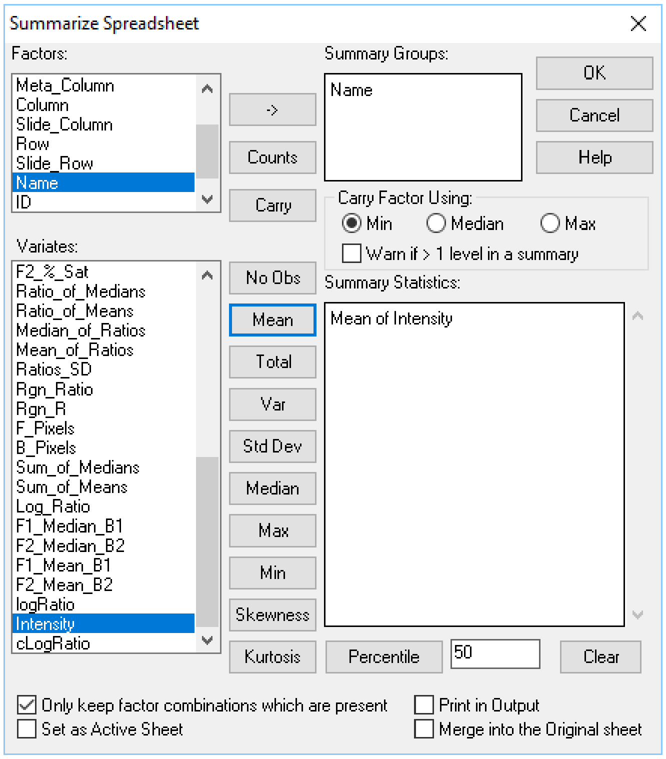
Available Data:
PVal[2]
CEst[1]
SEC[1]
CTVal[1]
CPVal[1]
m_Intensity
Mod_Cont
Mod_SD
Mod_Pr

Y variate: CPVal[1]

X variate: CEst[1]

Shade by: m_Intensity

Y Transformation: -Log10(Y)


Summarize Spreadsheet

Factors:
Meta_Column
Column
Slide_Column
Row
Slide_Row
Name
ID

Summary Groups: Name

Variates:
F2_%_Sat
Ratio_of_Medians
Ratio_of_Means
Median_of_Ratios
Mean_of_Ratios
Ratios_SD
Rgn_Ratio
Rgn_R
F_Pixels
B_Pixels
Sum_of_Medians
Sum_of_Means
Log_Ratio
F1_Median_B1
F2_Median_B2
F1_Mean_B1
F2_Mean_B2
logRatio
Intensity
cLogRatio

Carry Factor Using: ● Min ○ Median ○ Max

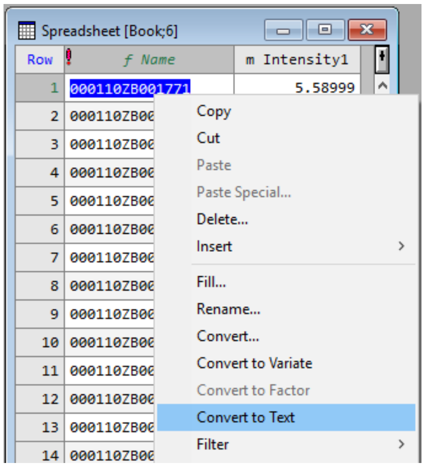
□ Warn if > 1 level in a summary

Summary Statistics: Mean of Intensity

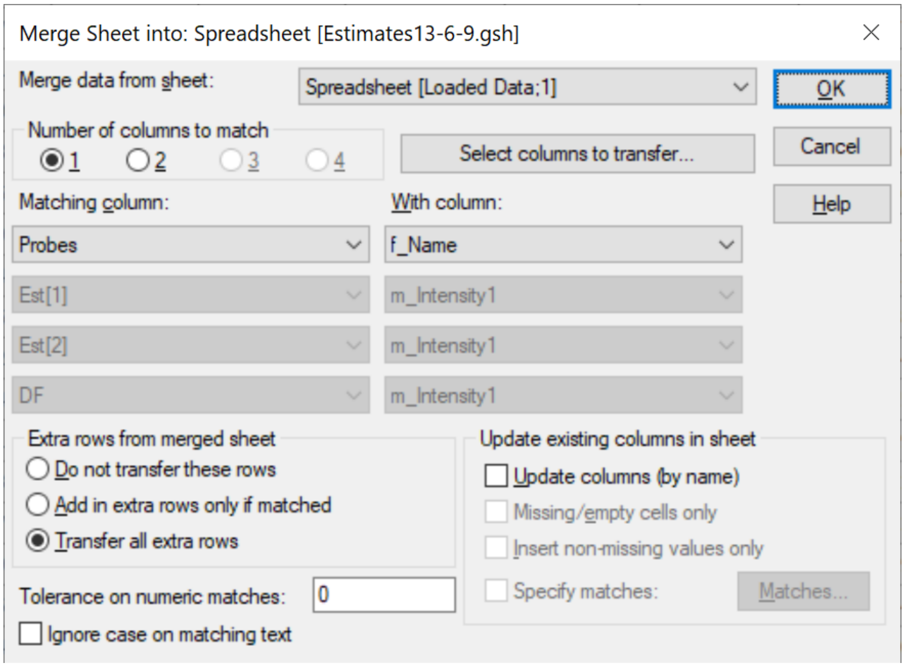
Percentile 50

☑ Only keep factor combinations which are present
□ Set as Active Sheet
□ Print in Output
□ Merge into the Original sheet

To merge the mean intensities with the Estimate results, the column f_Name must be converted to a text to match the type of the column Probes in the Estimates spreadsheet. To do this, right click on the column and select the popup menu Convert to Text. Now go to the Estimates spreadsheet window and use the Spread | Manipulate | Merge menu which opens the dialog below. The options are set correctly by default, so just click **OK** to merge the m_Intensity column.


Spreadsheet [Book;6]

Row	f Name	m Intensity1
1	000110ZB001771	5.58999
2	000110ZB00	
3	000110ZB00	
4	000110ZB00	
5	000110ZB00	
6	000110ZB00	
7	000110ZB00	
8	000110ZB00	
9	000110ZB00	
10	000110ZB00	
11	000110ZB00	
12	000110ZB00	
13	000110ZB00	
14	000110ZB00	


Merge Sheet into: Spreadsheet [Estimates13-6-9.gsh]

Merge data from sheet: Spreadsheet [Loaded Data;1]

Number of columns to match: ● 1 ○ 2 ○ 3 ○ 4

Matching column: Probes — With column: f_Name

Est[1] — m_Intensity1

Est[2] — m_Intensity1

DF — m_Intensity1

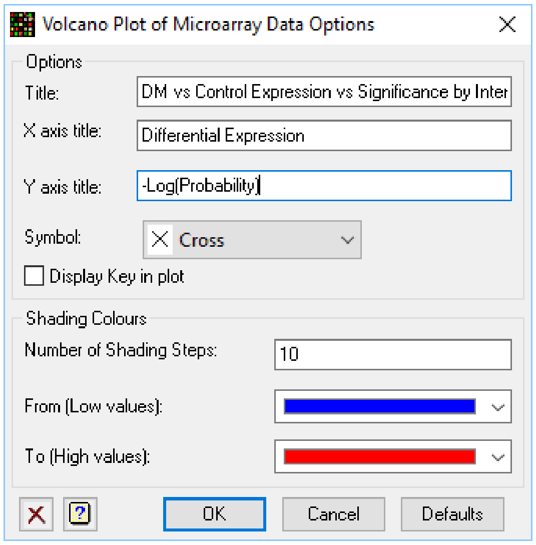
Extra rows from merged sheet:
○ Do not transfer these rows
○ Add in extra rows only if matched
● Transfer all extra rows

Update existing columns in sheet:
□ Update columns (by name)
□ Missing/empty cells only
□ Insert non-missing values only
□ Specify matches:

Tolerance on numeric matches: 0

□ Ignore case on matching text

Now, complete the Volcano Plot menu as shown on the previous page. The Options button allows you to set the titles, symbols and colours used on the plot. The options shown below left were used to produce the following graph (below right).


Volcano Plot of Microarray Data Options

Options

Title: DM vs Control Expression vs Significance by Inter

X axis title: Differential Expression

Y axis title: -Log(Probability)

Symbol: × Cross

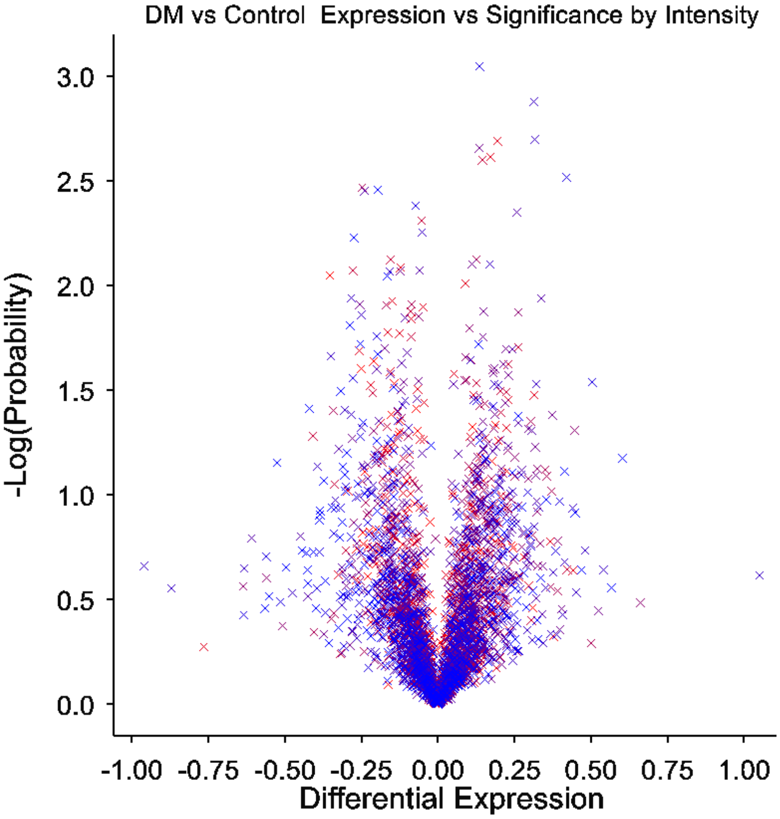
□ Display Key in plot

Shading Colours

Number of Shading Steps: 10

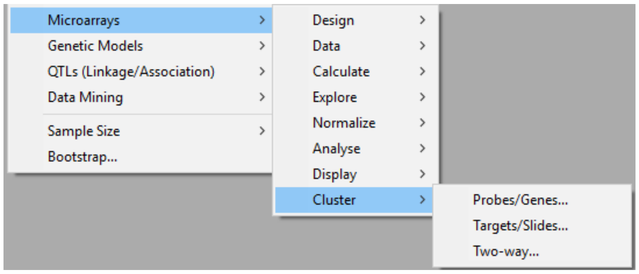
From (Low values): 🟦

To (High values): 🟥



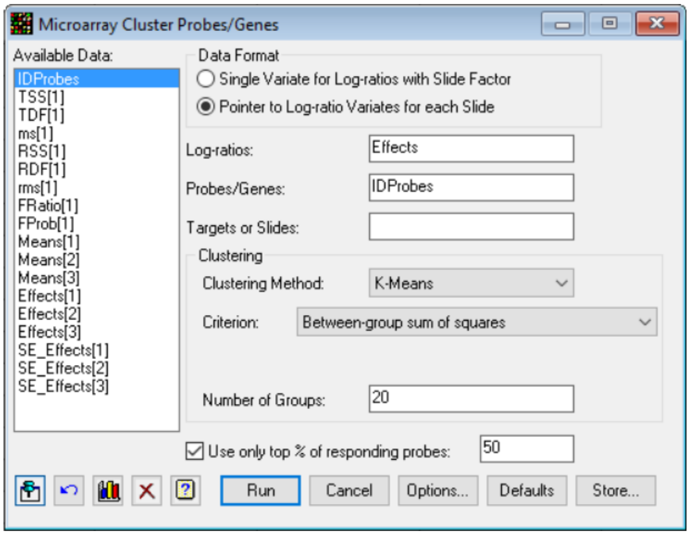
Cluster microarray probes or targets

This menu allows you to group similar probes or targets together through the technique of cluster analysis. This can be used to check that the similarity between slides reflects that expected from the experimental design, and to see which groups of genes/probes behave in a similar manner between targets. A two-way clustering using both slides and targets can be used to produce a 'heat map' of the probe x target matrix.



Cluster Probes/Genes

This menu can be used to group together probes or genes that behave in a similar manner over the various treatments in the experiment. As we often have a very large number of probes on a slide, there is an option to restrict the clustering to probes with the largest levels of differential expression. The data can be put into the menu in two formats, either as a pointer (see Help | Genstat Guides | Introduction, section 10.7 on page 328) to the results by slide or target or using single variate with all a factor indexing the various slides or targets.



Available Data: IDProbes, TSS[1], TDF[1], ms[1], RSS[1], RDF[1], rms[1], FRatio[1], FProb[1], Means[1], Means[2], Means[3], Effects[1], Effects[2], Effects[3], SE_Effects[1], SE_Effects[2], SE_Effects[3]

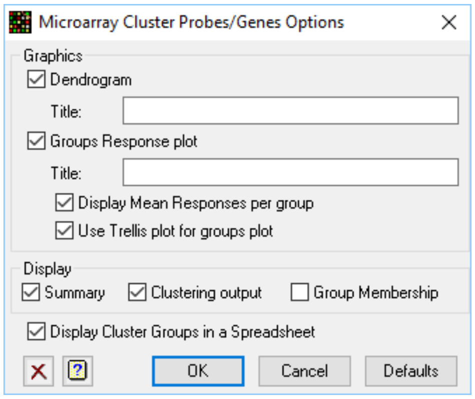
Data Format
○ Single Variate for Log-ratios with Slide Factor
● Pointer to Log-ratio Variates for each Slide

Log-ratios: Effects
Probes/Genes: IDProbes
Targets or Slides:

Clustering
Clustering Method: K-Means
Criterion: Between-group sum of squares
Number of Groups: 20

☑ Use only top % of responding probes: 50

For example, using the probes effects from the ANOVA of the nine Affymetrix slides saved in "Hyb-ANOVA.gwb", we can look to see which probes have similar responses over the three treatments. Opening the Stats | Microarray | Cluster | Probes/Genes menu, we get the dialog to the right. As the effects are in three columns pointed to by the Effects pointer, we select the **Pointer Data Format** as shown. We then select Effects as the **Log-Ratios** and enter IDProbes as the **Probes/Genes**. With a pointer, the **Targets or Slides** field can be left blank. The cluster method we use is **K means** and we specify the number of groups to cluster into as 20, and that we only want to cluster the top 50% of probes. Note: The K-Means algorithm is much more space efficient with large number of probes, as a hierarchical cluster analysis must calculate a full *n* x *n* similarity matrix, where *n* is the number of probes being clustered.

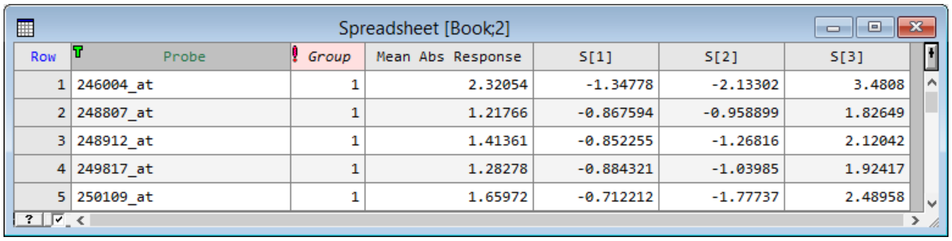


Graphics
☑ Dendrogram
Title:
☑ Groups Response plot
Title:
☑ Display Mean Responses per group
☑ Use Trellis plot for groups plot

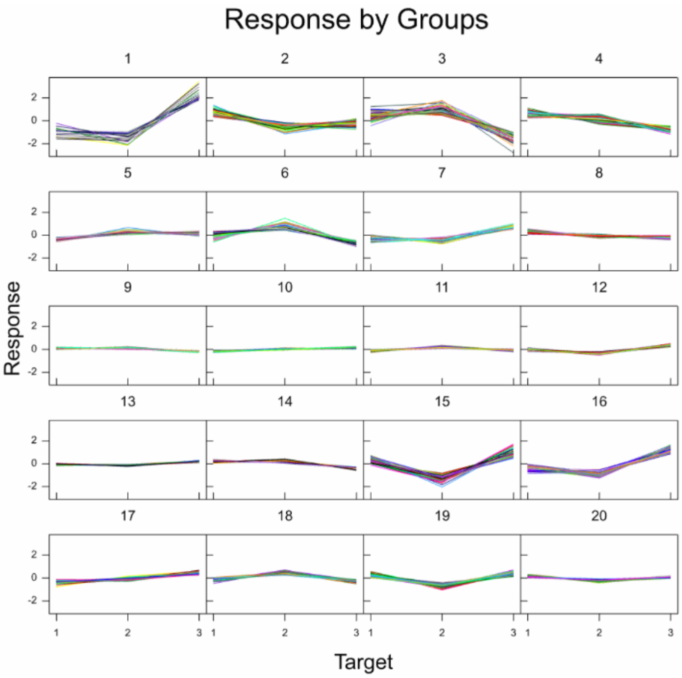
Display
☑ Summary ☑ Clustering output ☐ Group Membership

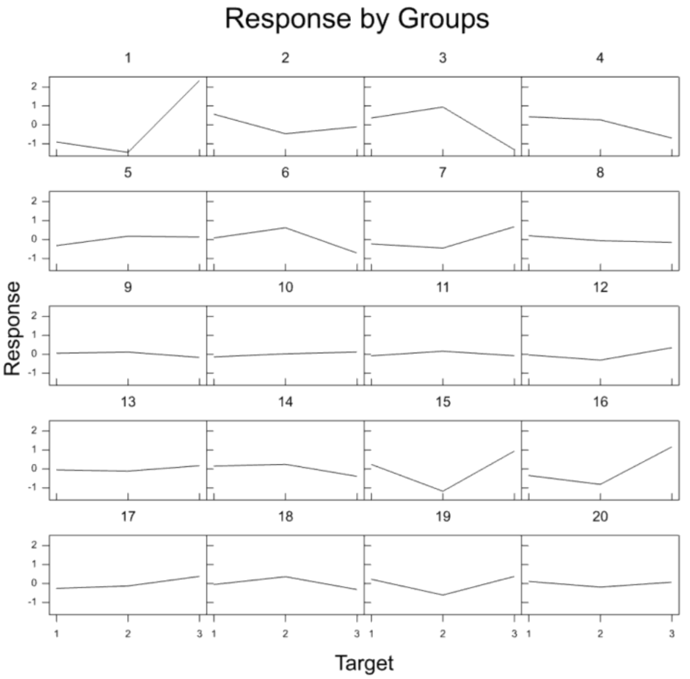
☑ Display Cluster Groups in a Spreadsheet

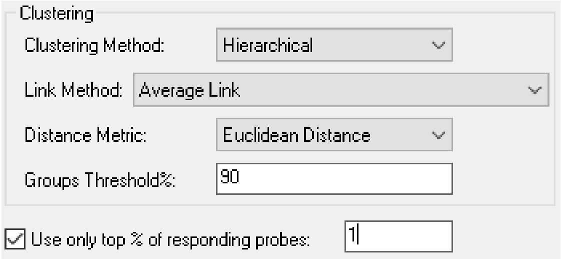
The Options dialog, shown to the right, controls the graphs and output from the clustering. If you use a Trellis plot with large numbers of probes, then the graph size in memory is often too large for most computers. Selecting the option **Display Mean Response per group** only shows the average over the probes in the groups defined by the clustering, considerably reducing the memory required to display the graph. The two graphs with and without this option are shown on the next page. The option **Display Cluster Groups in a Spreadsheet** produces the following spreadsheet that gives the probe groupings and the effects in columns S[1]…S[3] (one column for each treatment).



Row	Probe	Group	Mean Abs Response	S[1]	S[2]	S[3]
1	246004_at	1	2.32054	-1.34778	-2.13302	3.4808
2	248807_at	1	1.21766	-0.867594	-0.958899	1.82649
3	248912_at	1	1.41361	-0.852255	-1.26816	2.12042
4	249817_at	1	1.28278	-0.884321	-1.03985	1.92417
5	250109_at	1	1.65972	-0.712212	-1.77737	2.48958







Clustering

Clustering Method: Hierarchical

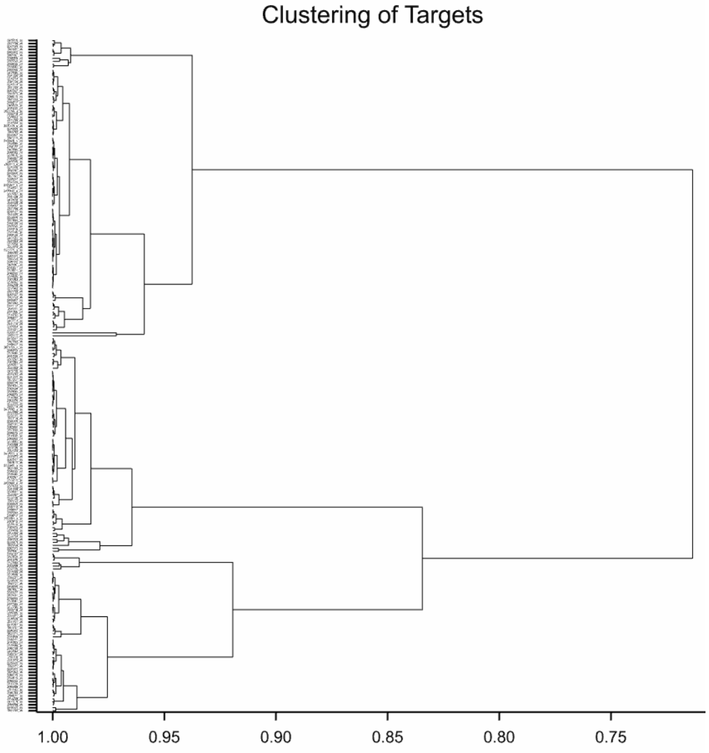
Link Method: Average Link

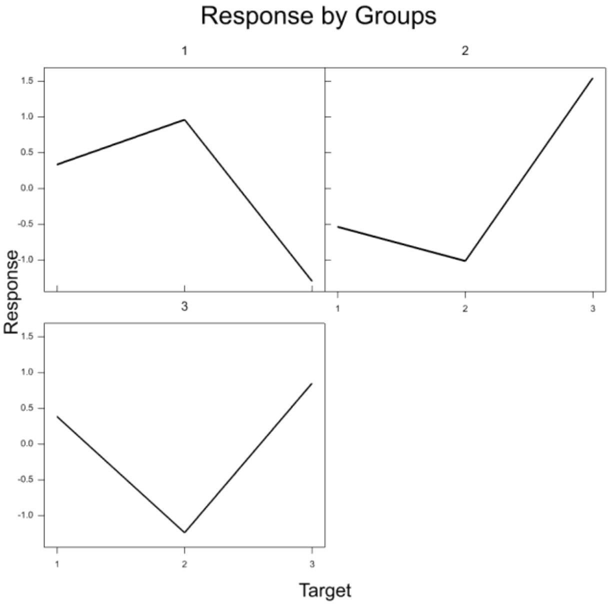
Distance Metric: Euclidean Distance

Groups Threshold%: 90

☑ Use only top % of responding probes: 1

If you use the hierarchical cluster algorithm (selected from the list under Clustering Method), with an average link, using only the top 1% of probes (settings shown below right), you will end up with the dendrogram on the left below. As you can see, it is difficult to read the probes with so many clustered together, even using only the top 1%, but the dendrogram does indicate that there are 3 quite distinct groupings of the probes. If you specify a Groups Threshold (see dialog below) of 90% (i.e. cutting through the dendrogram at a similarity of 0.9), you can see the mean responses of these 3 groups (as in the plot bottom right which was made using the trellis option): group 1 is high on Standard, group 2 low on Line-2 and group 3 low on Standard.

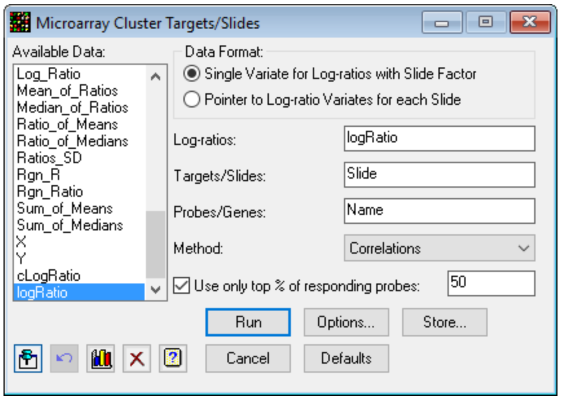




Cluster targets/slides

This menu lets you examine how the slides are related to each other in terms of having the same patterns of responses over the probes. The measurement used to assess similarity in a two colour experiment is usually the log-ratio, but this menu can also be used with the intensities from a single colour slide. Using the data from the "Data13-6-9.gwb" file, we can look at the similarities using the Correlations between slides (selected from the Method list) using the menu as shown to the right. This generates the dendrogram as shown below right (after some editing in the graphics editor). This shows slides 13-8 and 13-9 are most similar and 13-6 is most dissimilar. Of course the direction of the dye swaps should be taken into account, and the signs of the log-ratio could be changed on the slides which have a standard treatment on green as opposed to which have this treatment on red.

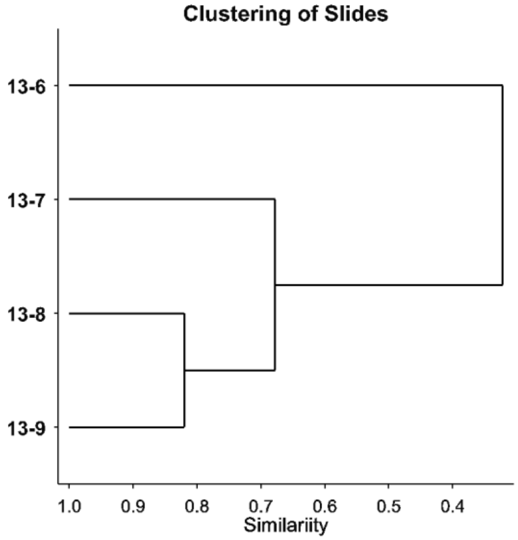
To look at the 9 Affymetrix slides in "Hyb-Expressions.gsh", we could use Expression in the Log-Ratios field with all the probes and obtain the following dendrogram (below left). When this is compared to the treatments on the slides it can be seen that the slide with the same treatments cluster together, and Line 1 and Line 2 are the most similar treatments.

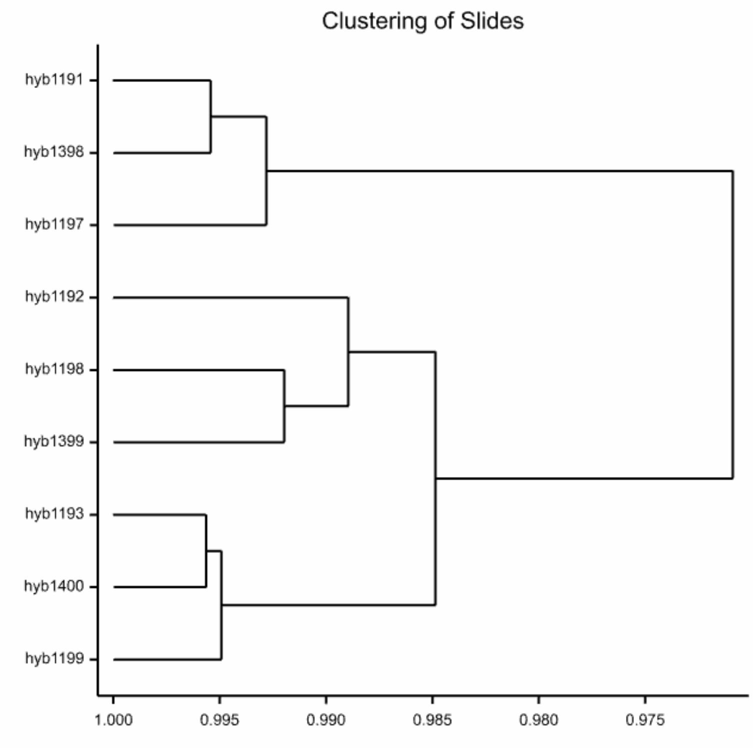

Microarray Cluster Targets/Slides

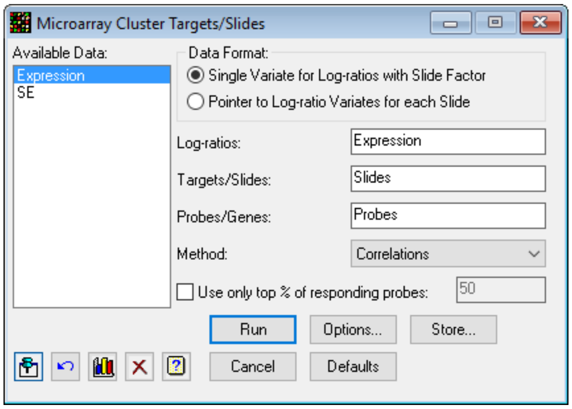
Available Data:

Data Format:
● Single Variate for Log-ratios with Slide Factor
○ Pointer to Log-ratio Variates for each Slide

Log-ratios: logRatio
Targets/Slides: Slide
Probes/Genes: Name
Method: Correlations
☑ Use only top % of responding probes: 50



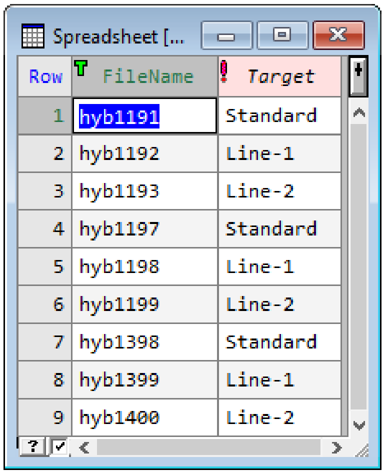



Microarray Cluster Targets/Slides

Available Data:

Data Format:
● Single Variate for Log-ratios with Slide Factor
○ Pointer to Log-ratio Variates for each Slide

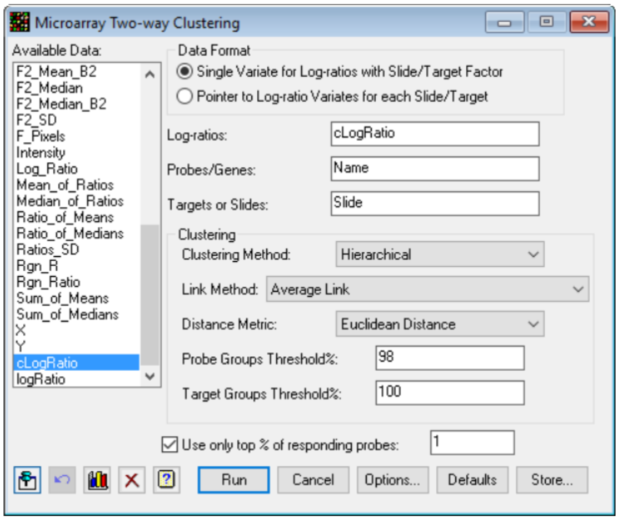
Log-ratios: Expression
Targets/Slides: Slides
Probes/Genes: Probes
Method: Correlations
□ Use only top % of responding probes: 50


Spreadsheet [...

Row	FileName	Target
1	hyb1191	Standard
2	hyb1192	Line-1
3	hyb1193	Line-2
4	hyb1197	Standard
5	hyb1198	Line-1
6	hyb1199	Line-2
7	hyb1398	Standard
8	hyb1399	Line-1
9	hyb1400	Line-2

Two-way clustering

This menu combines the previous two menus to allow you to jointly cluster on both probes/genes and targets/slides. A map of the two-way array of expression can then be displayed as a shade plot, which shows the pattern of differential expression. Using the example data saved in "Data13-6-9.gwb", open the Stats | Microarray | Cluster | Two-way menu to get the dialog to the right. Use the corrected log-ratios from the column cLogRatio, and enter the Probe and Slide information from the columns Name and Slide respectively. Completing the dialog as shown. Using a hierarchical clustering, only the top 1% of probes with the groups thresholds set to 98% for Probes and 100% for Targets (i.e. do not group any slides together), and setting the options as shown to the right, we obtain the dendrograms shown below (one for probes (below left) and another for slides (below right)) and the shade plot showing the group means by slides (bottom right).


Microarray Two-way Clustering

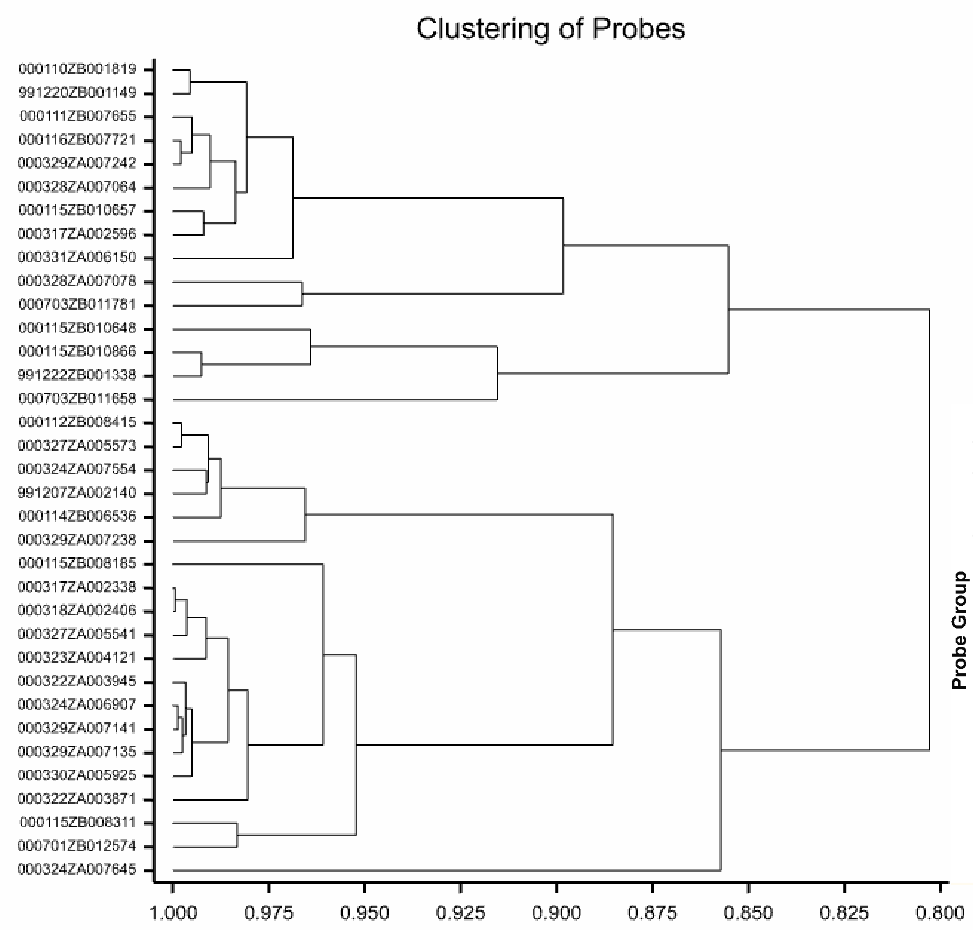
Available Data:
F2_Mean_B2
F2_Median
F2_Median_B2
F2_SD
F_Pixels
Intensity
Log_Ratio
Mean_of_Ratios
Median_of_Ratios
Ratio_of_Means
Ratio_of_Medians
Ratios_SD
Rgn_R
Rgn_Ratio
Sum_of_Means
Sum_of_Medians
X
Y
cLogRatio
logRatio

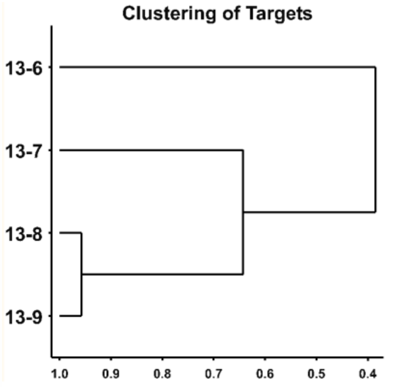
Data Format
● Single Variate for Log-ratios with Slide/Target Factor
○ Pointer to Log-ratio Variates for each Slide/Target

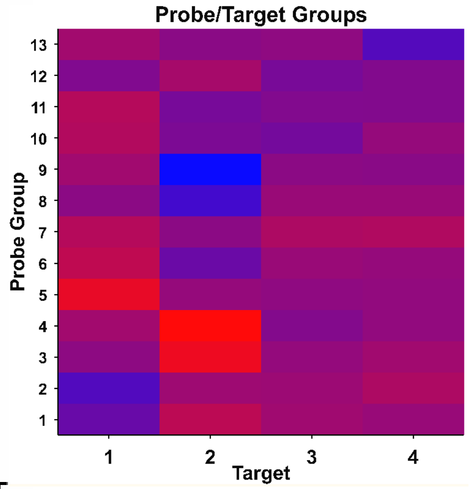
Log-ratios: cLogRatio
Probes/Genes: Name
Targets or Slides: Slide

Clustering
Clustering Method: Hierarchical
Link Method: Average Link
Distance Metric: Euclidean Distance
Probe Groups Threshold%: 98
Target Groups Threshold%: 100

☑ Use only top % of responding probes: 1

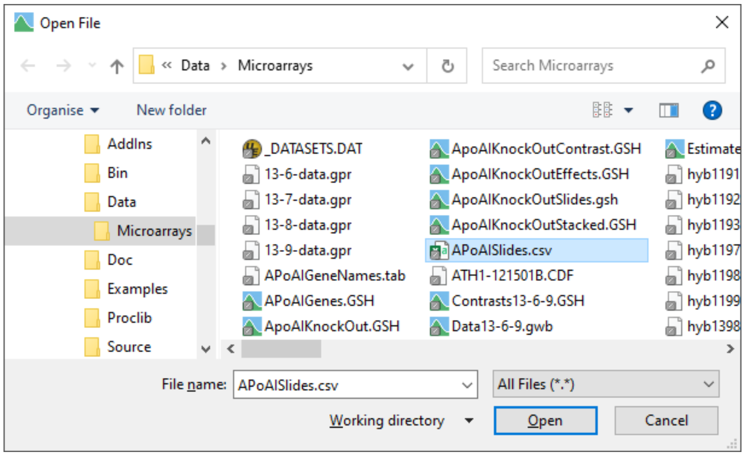






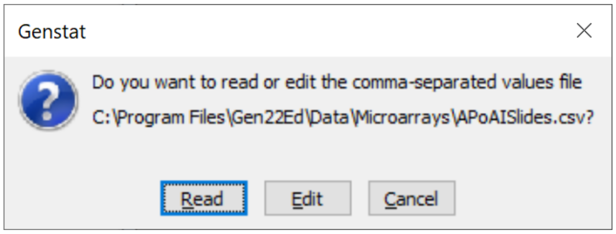
Two-channel microarray example

The following example shows how to analyse data from a two-channel microarray experiment. The data in this experiment are from a mouse knock out experiment with 6384 genes per slide. There were 16 slides, 8 control mice and 8 knockout mice all on the red dye compared to a standard reference on the green dye. Note that this design is not dye balanced, as there are no dye swaps, as the reference is always on red. The data are stored in the file "ApoAISlides.csv" that can be found in the Microarrays folder (as explained on page 2). The file can be opened in Genstat by selecting Open from the File menu and then navigating to and selecting the file name.



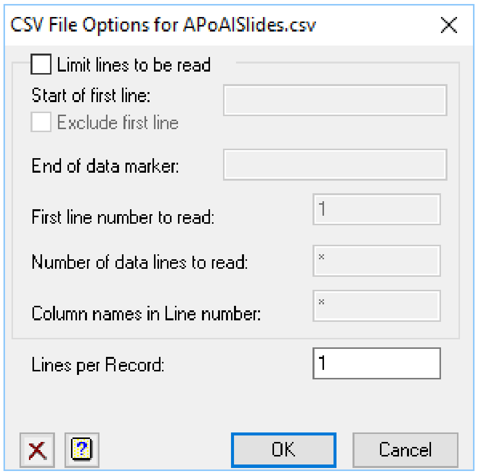
File name: APoAISlides.csv
All Files (*.*)

When a CSV file is opened in Genstat, you have the option of opening it into a text window, or into a spreadsheet. In this example, the data are to be opened into a spreadsheet: this can be done by clicking on the **Read** button as shown right.



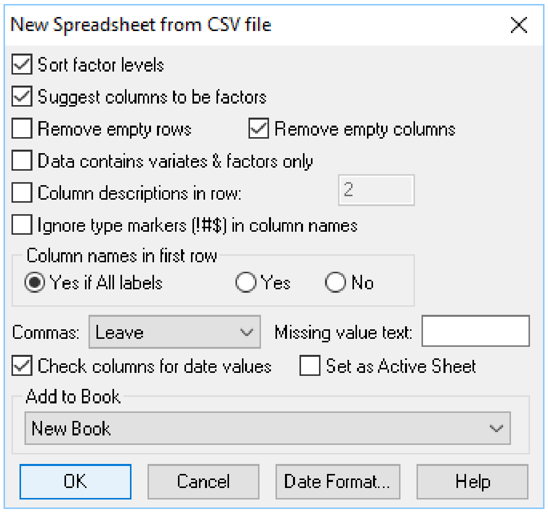
Do you want to read or edit the comma-separated values file C:\Program Files\Gen22Ed\Data\Microarrays\APoAISlides.csv?

On opening CSV files, you are prompted with two dialogs where additional options can be specified to control how the data are opened. The first dialog (as shown right) has options for controlling which rows of the data are to be opened. For this example, the whole file is read by clicking on the **OK** button.



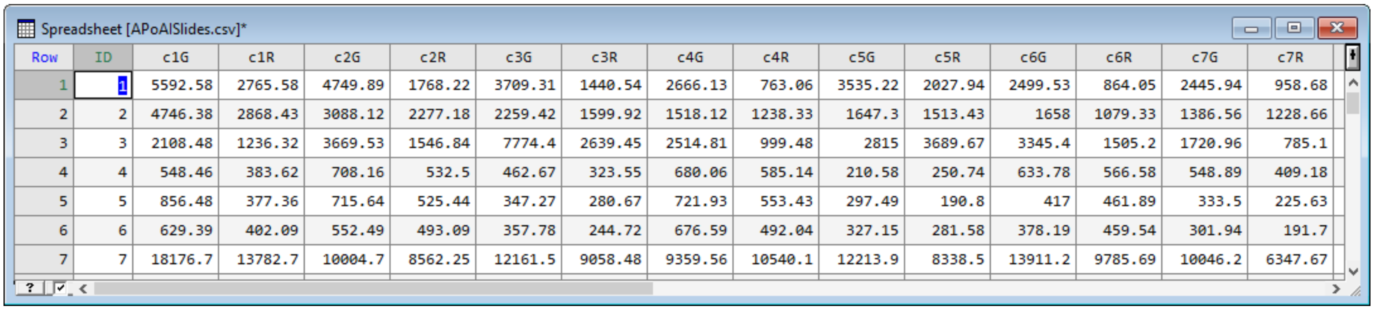
□ Limit lines to be read
Start of first line:
□ Exclude first line
End of data marker:
First line number to read: 1
Number of data lines to read: *
Column names in Line number: *
Lines per Record: 1

The second dialog contains further options for controlling how the data are to be opened including data type conversion and location of column names. For the example, the default settings can be used by clicking on the **OK** button.



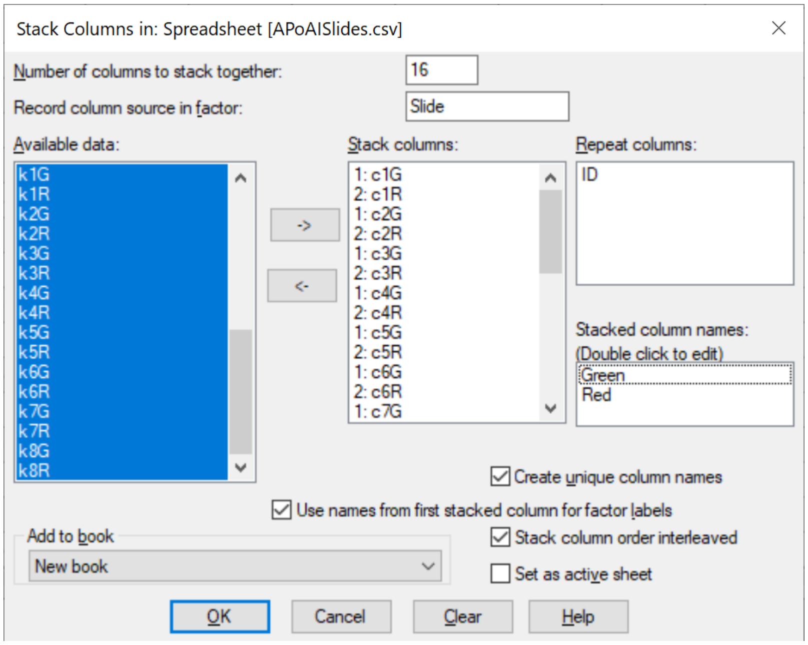
☑ Sort factor levels
☑ Suggest columns to be factors
□ Remove empty rows ☑ Remove empty columns
□ Data contains variates & factors only
□ Column descriptions in row: 2
□ Ignore type markers (!#$) in column names
Column names in first row
● Yes if All labels ○ Yes ○ No
Commas: Leave Missing value text:
☑ Check columns for date values □ Set as Active Sheet
Add to Book
New Book

Opening the file should result in the following spreadsheet:



Spreadsheet [APoAISlides.csv]*

Row	ID	c1G	c1R	c2G	c2R	c3G	c3R	c4G	c4R	c5G	c5R	c6G	c6R	c7G	c7R
1	1	5592.58	2765.58	4749.89	1768.22	3709.31	1440.54	2666.13	763.06	3535.22	2027.94	2499.53	864.05	2445.94	958.68
2	2	4746.38	2868.43	3088.12	2277.18	2259.42	1599.92	1518.12	1238.33	1647.3	1513.43	1658	1079.33	1386.56	1228.66
3	3	2108.48	1236.32	3669.53	1546.84	7774.4	2639.45	2514.81	999.48	2815	3689.67	3345.4	1505.2	1720.96	785.1
4	4	548.46	383.62	708.16	532.5	462.67	323.55	680.06	585.14	210.58	250.74	633.78	566.58	548.89	409.18
5	5	856.48	377.36	715.64	525.44	347.27	280.67	721.93	553.43	297.49	190.8	417	461.89	333.5	225.63
6	6	629.39	402.09	552.49	493.09	357.78	244.72	676.59	492.04	327.15	281.58	378.19	459.54	301.94	191.7
7	7	18176.7	13782.7	10004.7	8562.25	12161.5	9058.48	9359.56	10540.1	12213.9	8338.5	13911.2	9785.69	10046.2	6347.67



Stack Columns in: Spreadsheet [APoAISlides.csv]

Number of columns to stack together: 16

Record column source in factor: Slide

Available data: k1G, k1R, k2G, k2R, k3G, k3R, k4G, k4R, k5G, k5R, k6G, k6R, k7G, k7R, k8G, k8R

Stack columns: 1: c1G, 2: c1R, 1: c2G, 2: c2R, 1: c3G, 2: c3R, 1: c4G, 2: c4R, 1: c5G, 2: c5R, 1: c6G, 2: c6R, 1: c7G

Repeat columns: ID

Stacked column names: (Double click to edit) Green, Red

☑ Create unique column names

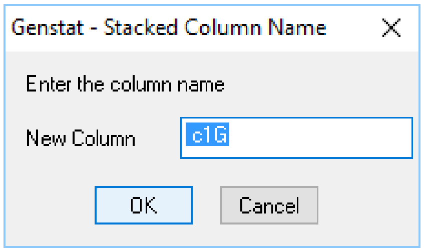
☑ Use names from first stacked column for factor labels

☑ Stack column order interleaved

☐ Set as active sheet

Add to book: New book

Within the spreadsheet, the data have two columns for each slide. To analyse the data, it needs to be in a stacked format where all the red values are within one column and all the green values are in another. To reorganise the data, the stack menu can be used. To open this dialog, select Stack from the Manipulate section of the Spread menu. The menu right shows the settings for stacking the columns together. There are 16 columns being stacked together. All the columns c1G-k8R are selected and added to the **Stacked Columns** list (using the ➔ button). Then tick the **Stack Column Order interleaved** box as we want alternating columns to go in the Green and Red columns. The name Slide has been entered for the factor to index the stacked columns. The column ID has been selected for the **Repeat Columns** list and the **Use names from First stacked column for Factor labels** has been ticked, so we get the column names as levels of the Slide factor.



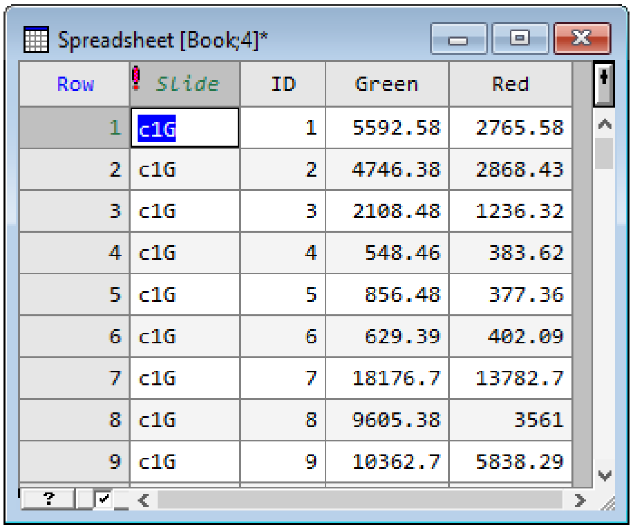
Genstat - Stacked Column Name

Enter the column name

New Column: c1G

Note that the stacked columns can be renamed to Green and Red by double clicking the old names in the **Stacked Columns Names** box and entering the new names in the rename dialog (see right).

Clicking **OK** on the stack menu should produce the spreadsheet to the right.



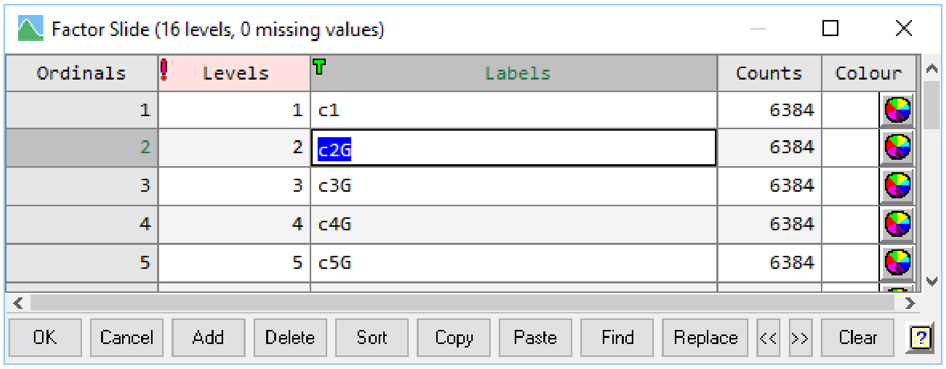
Spreadsheet [Book;4]*

Row	Slide	ID	Green	Red
1	c1G	1	5592.58	2765.58
2	c1G	2	4746.38	2868.43
3	c1G	3	2108.48	1236.32
4	c1G	4	548.46	383.62
5	c1G	5	856.48	377.36
6	c1G	6	629.39	402.09
7	c1G	7	18176.7	13782.7
8	c1G	8	9605.38	3561
9	c1G	9	10362.7	5838.29

The labels of the factor Slide have been created using the original column names. However, it may be preferable to change these labels to remove the ‘G’ to display the shorter labels c1...c8, k1...k8. The simplest way to do this is to select Edit Levels and Labels from the Factor item on the Spread menu or by clicking on the toolbar button.

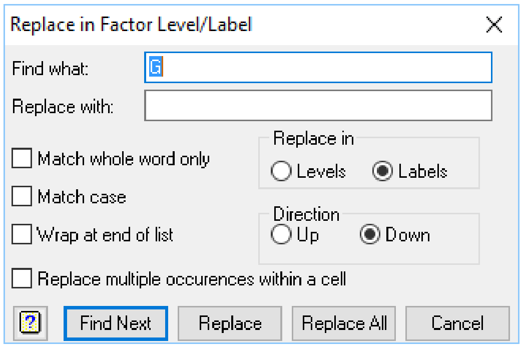
This will open the dialog shown right, where the ‘G’ can be removed from the labels by editing the appropriate cells.

A faster way of doing this is to use the Replace button and replace G with nothing:



Factor Slide (16 levels, 0 missing values)

Ordinals	Levels	Labels	Counts	Colour
1	1	c1	6384	●
2	2	c2G	6384	●
3	3	c3G	6384	●
4	4	c4G	6384	●
5	5	c5G	6384	●



Replace in Factor Level/Label

Find what: G

Replace with:

□ Match whole word only

□ Match case

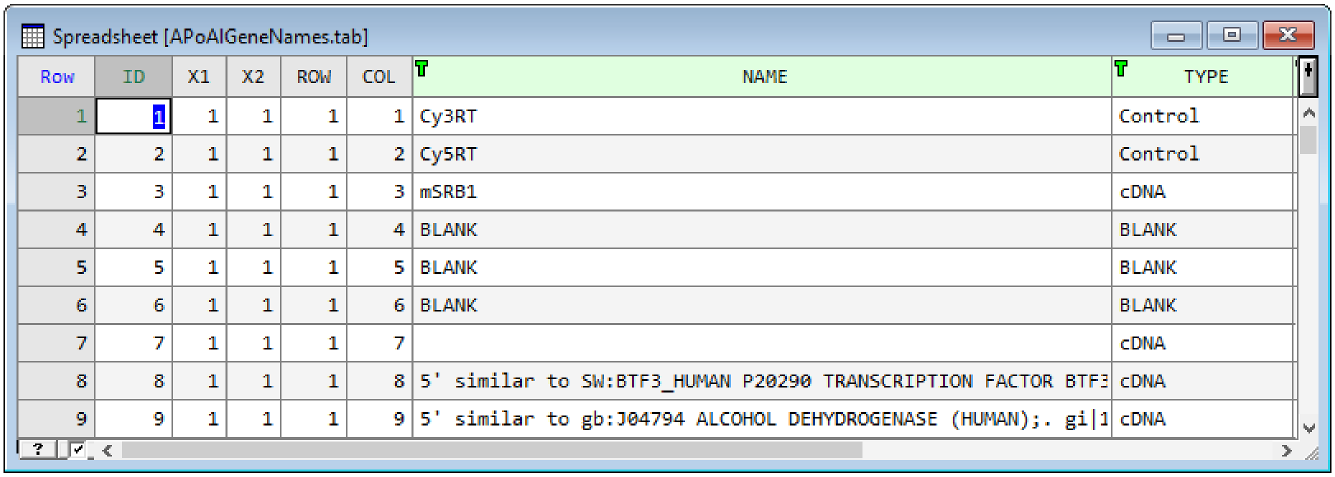
□ Wrap at end of list

Replace in: ○ Levels ● Labels

Direction: ○ Up ● Down

□ Replace multiple occurences within a cell

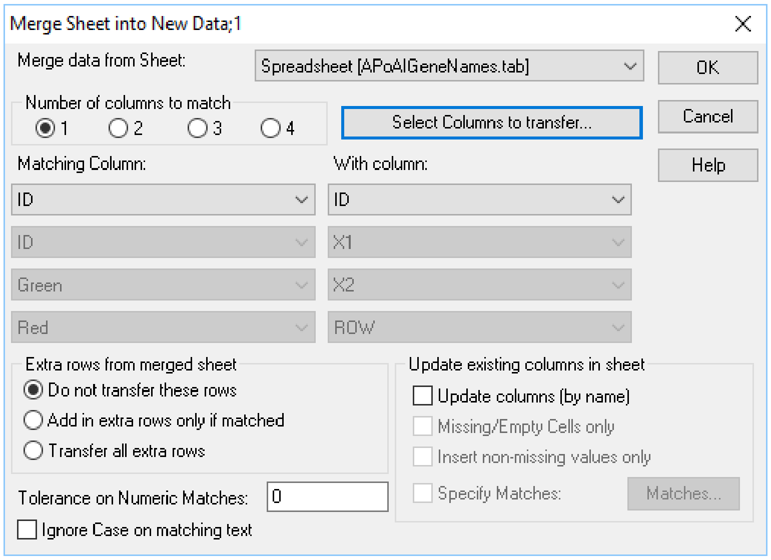
Additional information on the genes and layout of the slides is located within another file, “ApoAIGeneNames.tab”. The file can be opened using the Open item on the File menu, and should result in the following spreadsheet:



Spreadsheet [APoAIGeneNames.tab]

Row	ID	X1	X2	ROW	COL	NAME	TYPE
1	1	1	1	1	1	Cy3RT	Control
2	2	1	1	1	2	Cy5RT	Control
3	3	1	1	1	3	mSRB1	cDNA
4	4	1	1	1	4	BLANK	BLANK
5	5	1	1	1	5	BLANK	BLANK
6	6	1	1	1	6	BLANK	BLANK
7	7	1	1	1	7		cDNA
8	8	1	1	1	8	5' similar to SW:BTF3_HUMAN P20290 TRANSCRIPTION FACTOR BTF3	cDNA
9	9	1	1	1	9	5' similar to gb:J04794 ALCOHOL DEHYDROGENASE (HUMAN);. gi|1	cDNA

The information from the “ApoAIGeneNames.tab” data set needs to be merged into the stacked spreadsheet. To merge two spreadsheets, click on the spreadsheet that the data are to be merged into (in this case the stacked spreadsheet). Select Merge from the Manipulate item on the Spread menu, which should open the dialog on the right. The two spreadsheets are to be merged using the column ID to match columns between the spreadsheets.



Merge Sheet into New Data;1

Merge data from Sheet: Spreadsheet [APoAIGeneNames.tab]

Number of columns to match: ● 1 ○ 2 ○ 3 ○ 4

Matching Column:	With column:
ID	ID
ID	X1
Green	X2
Red	ROW

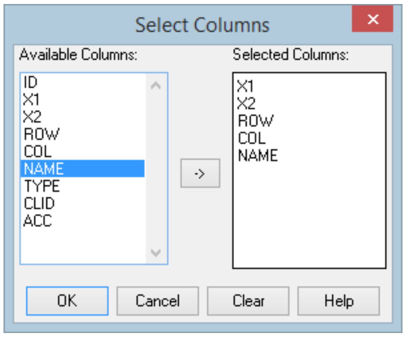
Extra rows from merged sheet
● Do not transfer these rows
○ Add in extra rows only if matched
○ Transfer all extra rows

Update existing columns in sheet
□ Update columns (by name)
□ Missing/Empty Cells only
□ Insert non-missing values only
□ Specify Matches:

Tolerance on Numeric Matches: 0

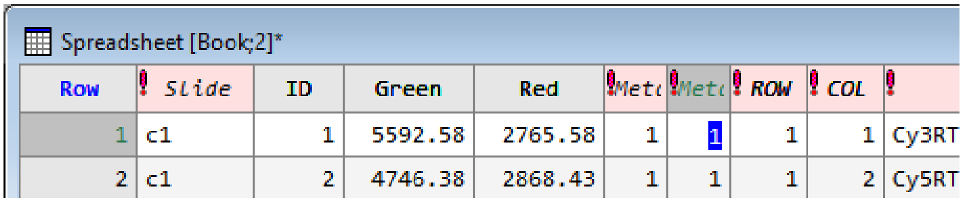
□ Ignore Case on matching text

The columns X1, X2, ROW, COL and NAME can be merged into the original spreadsheet by clicking on the **Select Columns to Transfer** button and then copying these names to the Selected Columns list and then clicking **OK**.



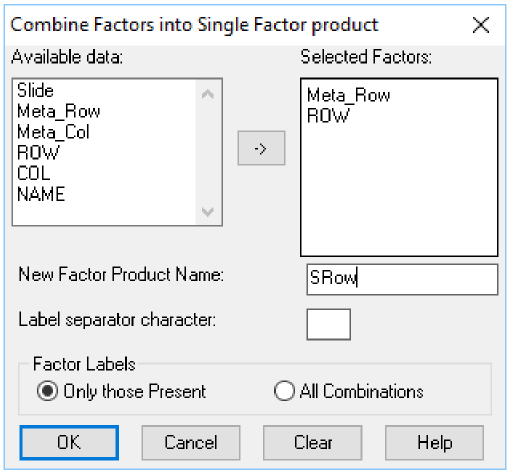
Selected Columns: X1, X2, ROW, COL, NAME

The column X1 is the row position of the pins down the slide, and X2 is the column position. These can be renamed to more the informative names Meta_Row and Meta_Col by clicking on the start of the column name (the cursor should change to a pencil when you hover at the start of the column name) and entering the new name. The spreadsheet columns which index the row and column layout of the pins (Meta_Row and Meta_Col), the rows and columns within pins (ROW and COL) and the Gene Names (NAME) should all be converted to factors. To convert columns to factors, right-click the mouse anywhere within the column to be converted, and then select the "Convert to Factor" from the pop-up menu. Once this has been done for each of the columns, the factor columns will be indicated by an exclamation mark at the start of the column name as seen to the right.


Spreadsheet [Book;2]*

Row	! Slide	ID	Green	Red	! Metc	! Metc	! ROW	! COL	!
1	c1	1	5592.58	2765.58	1	1	1	1	Cy3RT
2	c1	2	4746.38	2868.43	1	1	1	2	Cy5RT

The row and column positions across the whole slide are required for the analysis. These can be formed by using the combinations of Meta_Row with ROW and Meta_Col with COL respectively. To form the product of factors, select Product/Combine from the Factor item on the Spread menu. This opens the dialog shown right, in which the two factors Meta_Row and ROW have been selected, and the new name SRow has been entered for the product. Similarly, this dialog can be used to form the product of Meta_Col with COL with a new name SCol. Note that if the data are to be analysed using the normalization menu, the factors Meta_Row and Meta_Col will need to be combined to form a factor representing the pins.



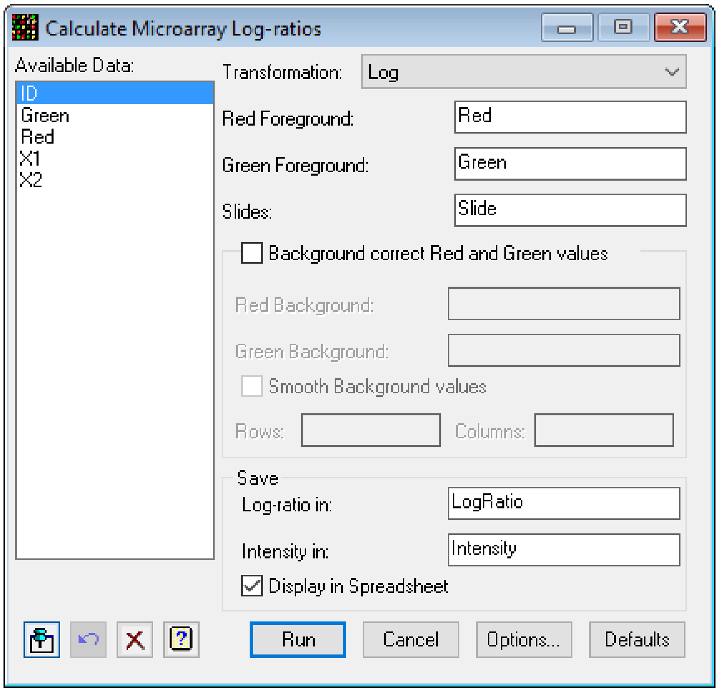
Selected Factors: Meta_Row, ROW

New Factor Product Name: SRow

Label separator character:

Factor Labels: ● Only those Present ○ All Combinations

To measure the level of differential expression between the two treatments on a slide the log-ratios can be calculated. To calculate the log-ratios select Log-ratios from the Calculate sub-menu from the Microarrays on the Stats menu. The menu right shows the settings that can be used to calculate the log-ratios for this data set.



Transformation: Log

Red Foreground: Red

Green Foreground: Green

Slides: Slide

□ Background correct Red and Green values

Red Background:

Green Background:

□ Smooth Background values

Rows: Columns:

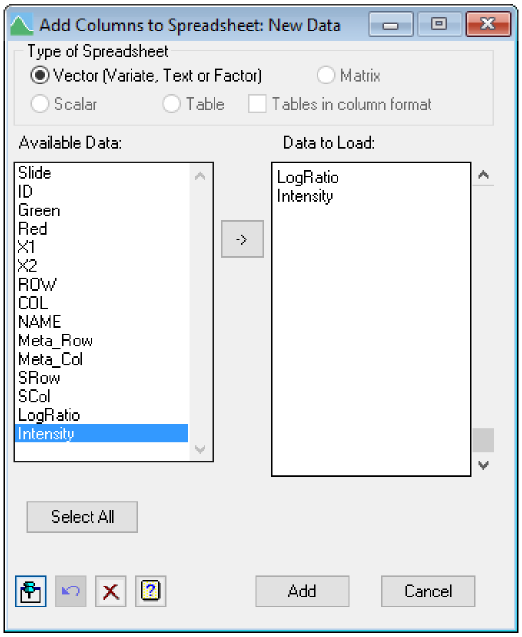
Save

Log-ratio in: LogRatio

Intensity in: Intensity

☑ Display in Spreadsheet

If the newly calculated log-ratio and intensity columns are not automatically added to your existing spreadsheet, you can append them by selecting the “Data in Genstat” item from Add on the Spread menu. In the corresponding dialog (right), select the two columns to be added to the spreadsheet and click on the **Add** button.



Add Columns to Spreadsheet: New Data

Type of Spreadsheet

● Vector (Variate, Text or Factor) ○ Matrix

○ Scalar ○ Table □ Tables in column format

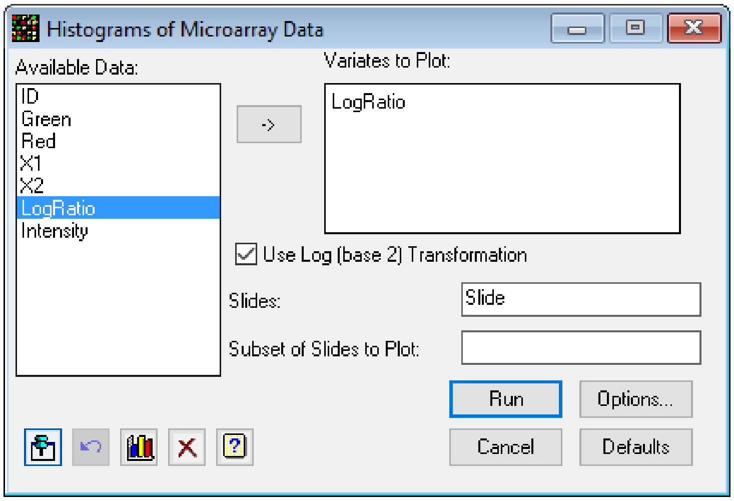
Available Data:

Slide
ID
Green
Red
X1
X2
ROW
COL
NAME
Meta_Row
Meta_Col
SRow
SCol
LogRatio
Intensity

Data to Load:

LogRatio
Intensity

The data on the slides can be explored by using the graphical menus available within the Explore sub-menu. For example, Histograms can be selected to produce histograms of the data. The following shows the settings for plotting a histogram of the log-ratios by slide.



Histograms of Microarray Data

Available Data:

ID
Green
Red
X1
X2
LogRatio
Intensity

Variates to Plot:

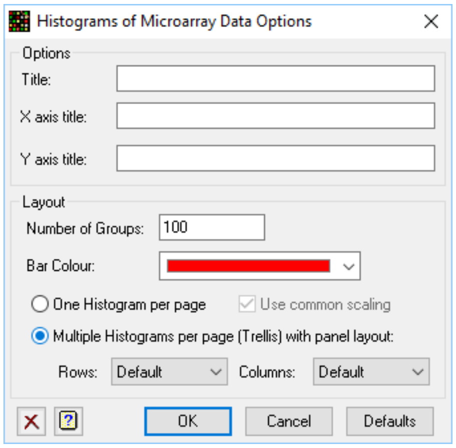
LogRatio

☑ Use Log (base 2) Transformation

Slides: Slide

Subset of Slides to Plot:

The options button on the menu should be used to set the Trellis options as below left, giving the graph below right.



Histograms of Microarray Data Options

Options

Title:

X axis title:

Y axis title:

Layout

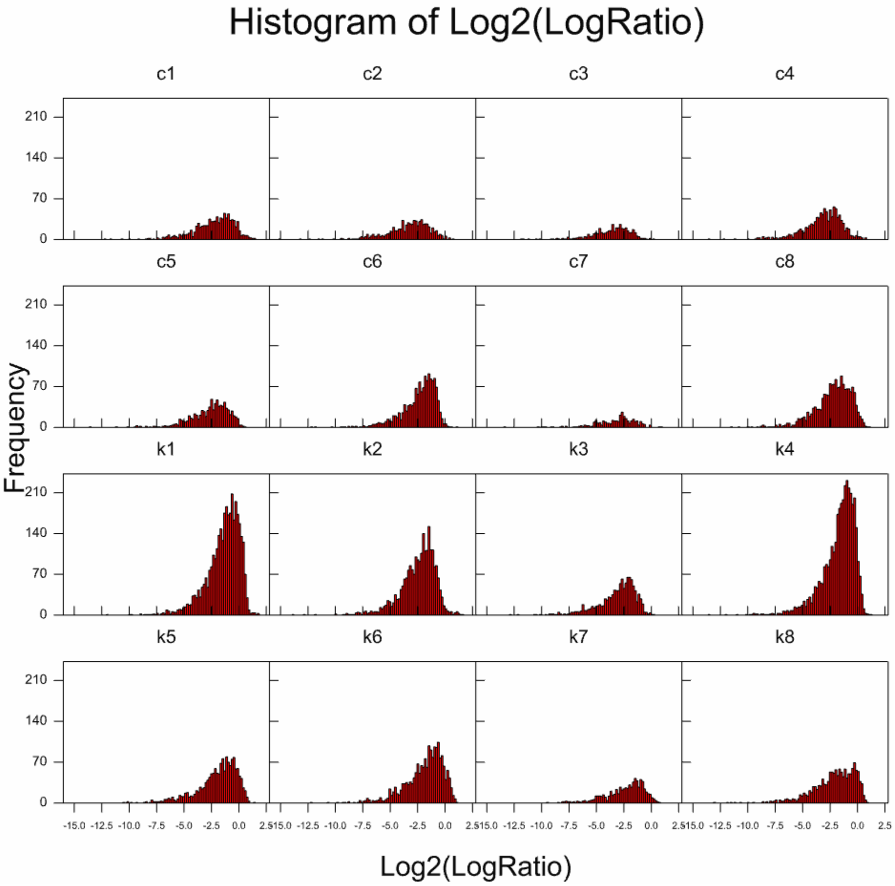
Number of Groups: 100

Bar Colour: 🟥

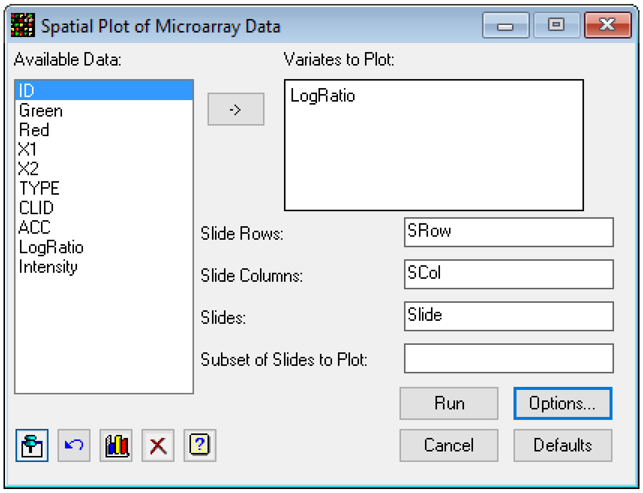
○ One Histogram per page ☑ Use common scaling

● Multiple Histograms per page (Trellis) with panel layout:

Rows: Default Columns: Default



The spatial variation across the slides can be examined by selecting the Spatial Plot item from the Explore sub-menu. The menu (right) shows the settings that can be used to produce a spatial plot for each slide. The spatial plot of the first slide is shown below.



Spatial Plot of Microarray Data

Available Data:
ID
Green
Red
X1
X2
TYPE
CLID
ACC
LogRatio
Intensity

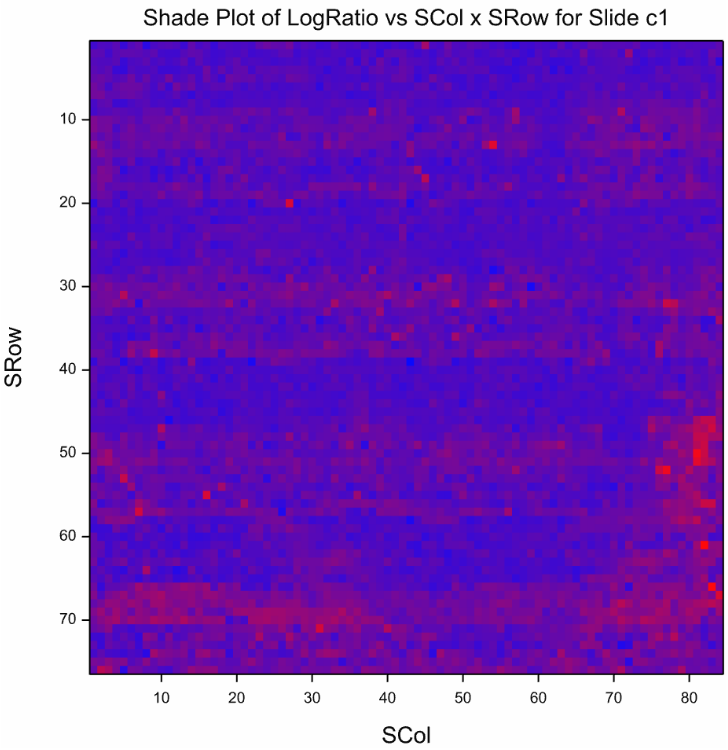
Variates to Plot: LogRatio

Slide Rows: SRow

Slide Columns: SCol

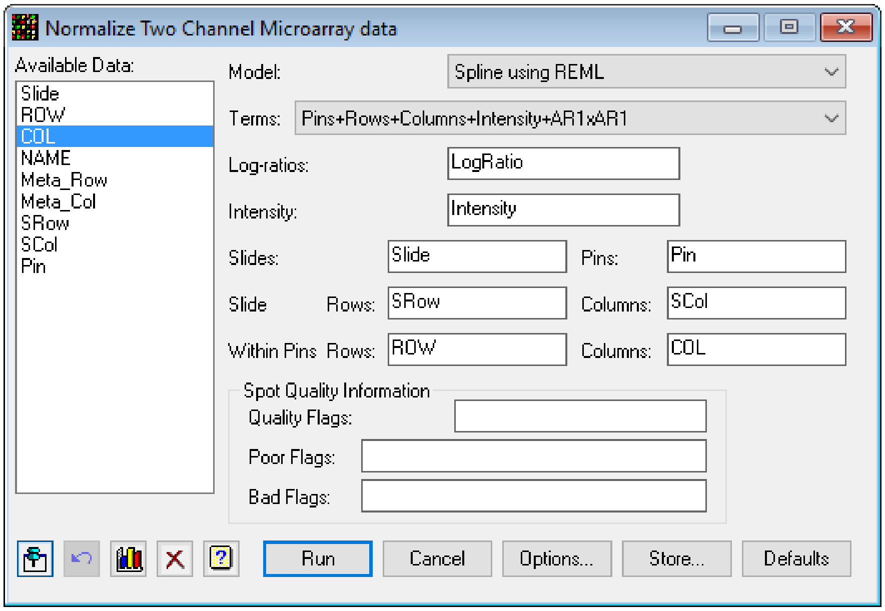
Slides: Slide

Subset of Slides to Plot:



Dye intensity and spatial effects (pins, rows, and columns) can be removed from the slides by using the Normalization menu. To open the normalization menu, select Two Channel from the Normalize item on the Microarrays menu. The menu below shows the settings that can be used to normalize the data.

Note that the option to include plots is available within the options for this menu. You will need to open the options dialog with Options button and set this as below right:



Normalize Two Channel Microarray data

Available Data:
Slide
ROW
COL
NAME
Meta_Row
Meta_Col
SRow
SCol
Pin

Model: Spline using REML

Terms: Pins+Rows+Columns+Intensity+AR1xAR1

Log-ratios: LogRatio

Intensity: Intensity

Slides: Slide   Pins: Pin

Slide Rows: SRow   Columns: SCol

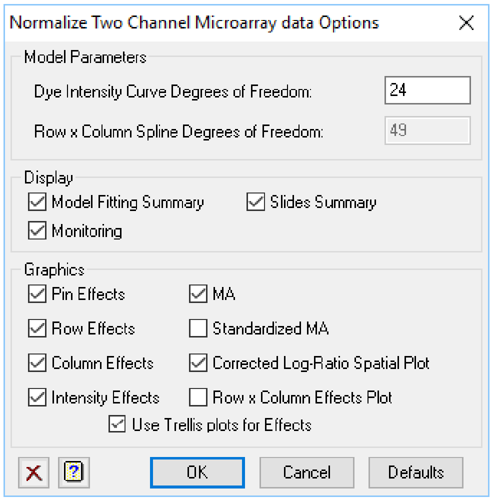
Within Pins Rows: ROW   Columns: COL

Spot Quality Information

Quality Flags:

Poor Flags:

Bad Flags:



Normalize Two Channel Microarray data Options

Model Parameters

Dye Intensity Curve Degrees of Freedom: 24

Row x Column Spline Degrees of Freedom: 49

Display

☑ Model Fitting Summary ☑ Slides Summary

☑ Monitoring

Graphics

☑ Pin Effects ☑ MA

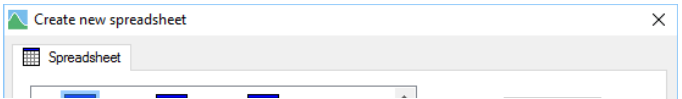
☑ Row Effects ☐ Standardized MA

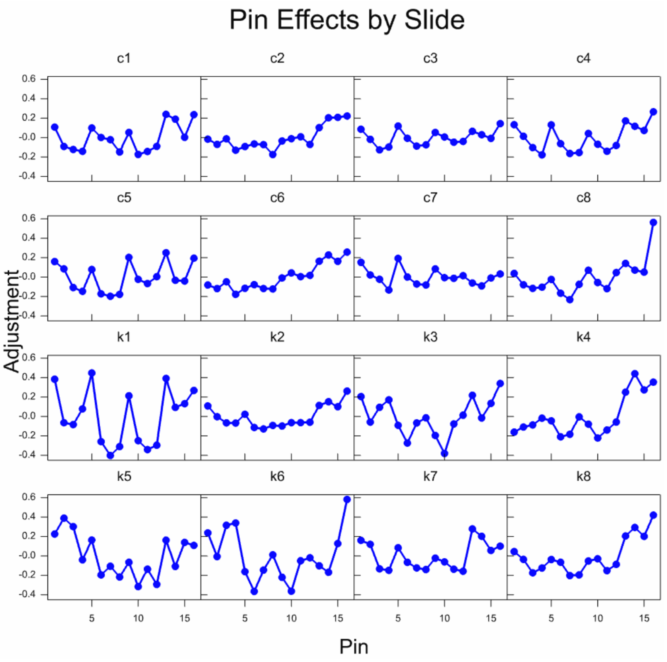
☑ Column Effects ☑ Corrected Log-Ratio Spatial Plot

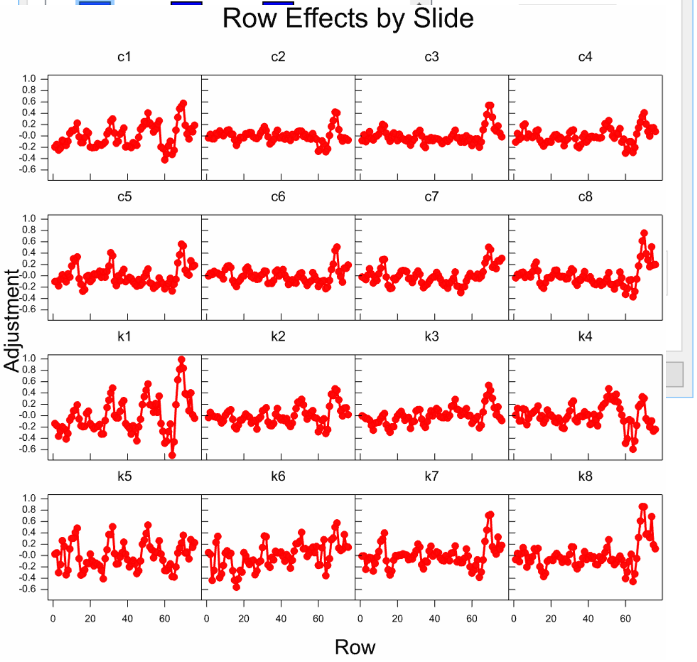
☑ Intensity Effects ☐ Row x Column Effects Plot

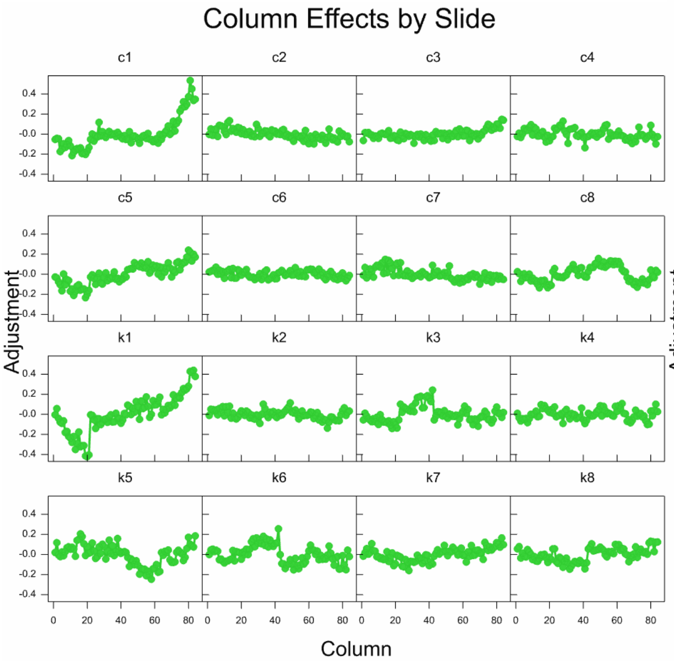
☑ Use Trellis plots for Effects

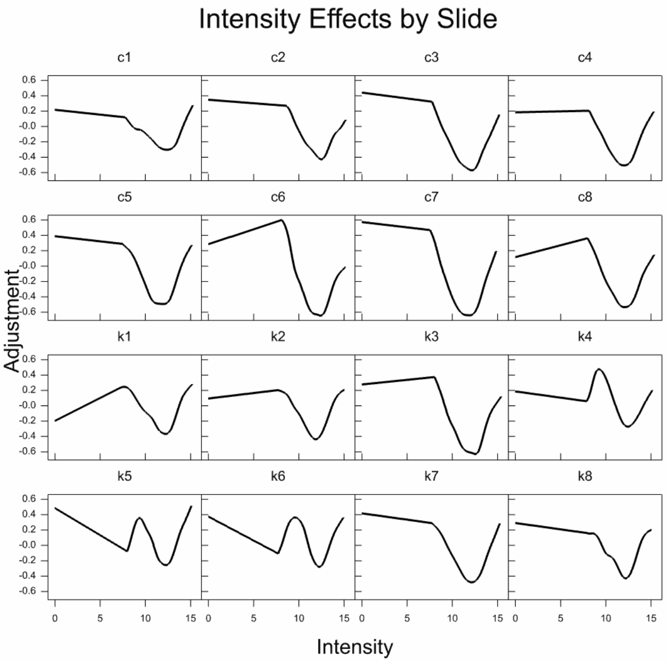
The resulting graphs display the effects that have been estimated:





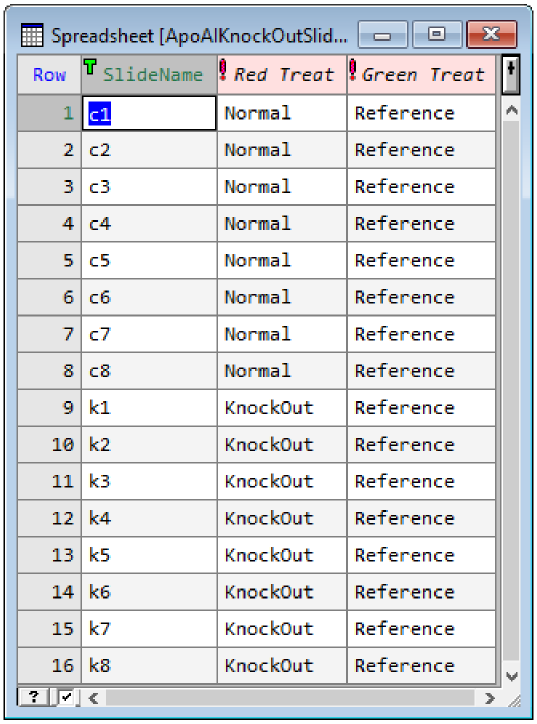






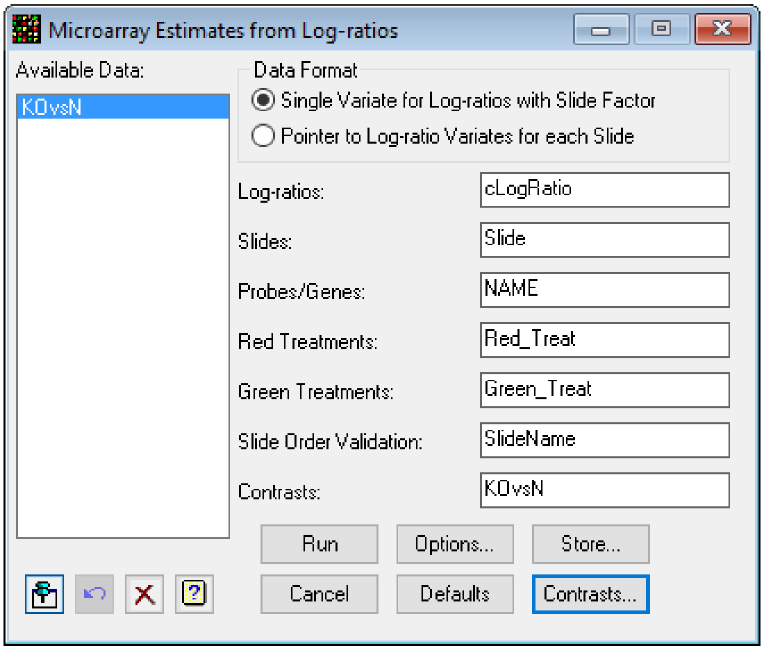
To analyse the results across the 16 slides, a small data set is required which provides the treatments applied to each slide. To do this an empty spreadsheet with 3 columns and 16 rows can be created using the menu shown right. This menu can be opened by selecting New from the File menu and then selecting the Spreadsheet tab.

The data can easily be entered into the empty spreadsheet. The spreadsheet below right shows the data that should be entered, with the three columns named SlideName, Red_Treat, and Green_Treat. Note that the columns Red_Treat and Green_Treat have been converted to factors. The factors columns Green_Treat and Red_Treat must contain the same set of factor levels or labels. In this example, the columns should be created such that they both have three labels (KnockOut, Normal and Reference).


Spreadsheet [ApoAIKnockOutSlid...

Row	SlideName	Red Treat	Green Treat
1	c1	Normal	Reference
2	c2	Normal	Reference
3	c3	Normal	Reference
4	c4	Normal	Reference
5	c5	Normal	Reference
6	c6	Normal	Reference
7	c7	Normal	Reference
8	c8	Normal	Reference
9	k1	KnockOut	Reference
10	k2	KnockOut	Reference
11	k3	KnockOut	Reference
12	k4	KnockOut	Reference
13	k5	KnockOut	Reference
14	k6	KnockOut	Reference
15	k7	KnockOut	Reference
16	k8	KnockOut	Reference

To analyse the data, select the Estimate Two Channel Effects item from Analyse on the Microarray menu. The picture below shows the resulting menu containing settings to run the analysis.

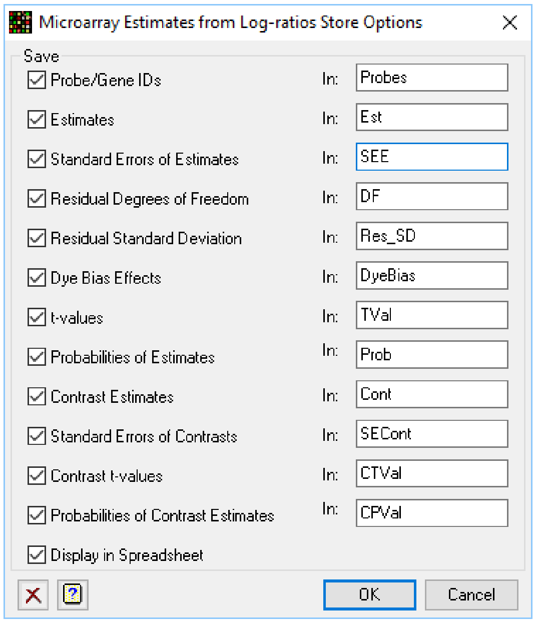

Microarray Estimates from Log-ratios

Available Data:
KOvsN

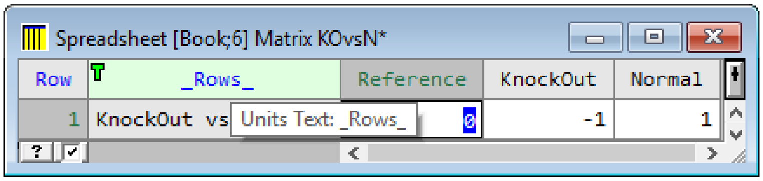
Data Format
● Single Variate for Log-ratios with Slide Factor
○ Pointer to Log-ratio Variates for each Slide

Log-ratios: cLogRatio
Slides: Slide
Probes/Genes: NAME
Red Treatments: Red_Treat
Green Treatments: Green_Treat
Slide Order Validation: SlideName
Contrasts: KOvsN

Results can be saved into data structures simultaneously with the analysis by clicking the **Store** button (before clicking **Run**) and specifying the names of the new structures. These structures can also be displayed into spreadsheets by selecting the **Display in Spreadsheet**. Note as no dye bias has been estimated, this will not appear in the resulting spreadsheet.


Microarray Estimates from Log-ratios Store Options

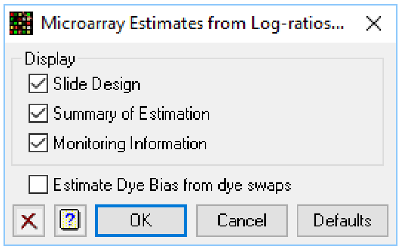
Save
☑ Probe/Gene IDs In: Probes
☑ Estimates In: Est
☑ Standard Errors of Estimates In: SEE
☑ Residual Degrees of Freedom In: DF
☑ Residual Standard Deviation In: Res_SD
☑ Dye Bias Effects In: DyeBias
☑ t-values In: TVal
☑ Probabilities of Estimates In: Prob
☑ Contrast Estimates In: Cont
☑ Standard Errors of Contrasts In: SECont
☑ Contrast t-values In: CTVal
☑ Probabilities of Contrast Estimates In: CPVal
☑ Display in Spreadsheet



Row	_Rows_	Reference	KnockOut	Normal
1	KnockOut vs Units Text: _Rows_	0	-1	1

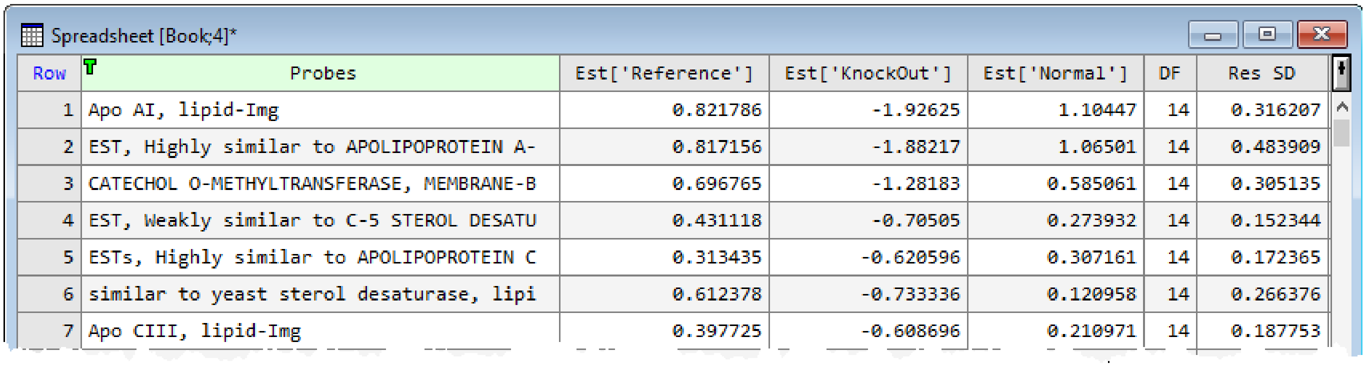
To estimate the difference between the control and the knockout treatments, a contrast can be defined by clicking on the **Contrasts** button. This prompts for the contrast matrix name (KOvsN) and the number of contrasts (1). Clicking **OK** pops-up a spreadsheet where the contrast matrix values can be supplied. The spreadsheet above shows a contrast for control versus knockout. Note the reference level is specified as zero as it is not used in this contrast.

For this example, the Estimate dye bias from dye swaps option is not required and should be removed by making sure the option is not selected within options (click on the **Options** button to view the menu options). Clicking on the **Run** button will run the analysis and display the results in a spreadsheet.



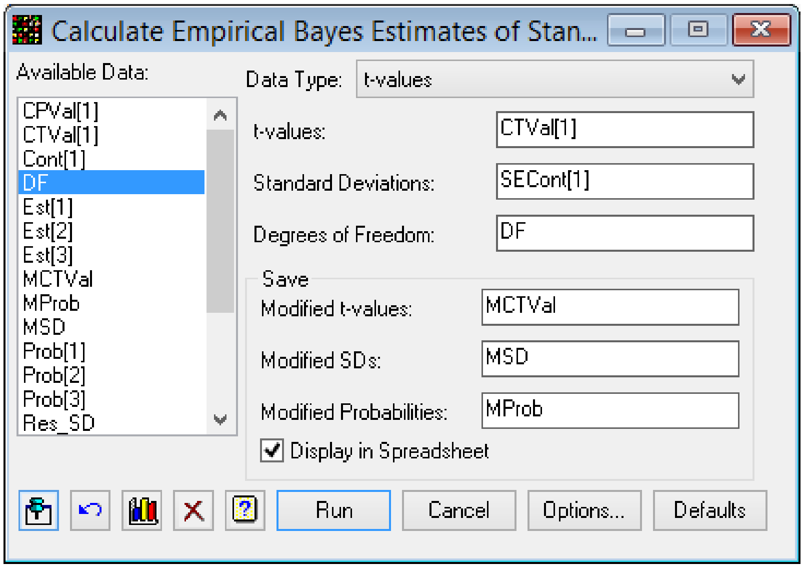
☑ Slide Design
☑ Summary of Estimation
☑ Monitoring Information
☐ Estimate Dye Bias from dye swaps

If the spreadsheet is sorted by the contrast (using the Spread | Sort menu item) it can be seen that the APO gene has the largest level of differential expression (as below).



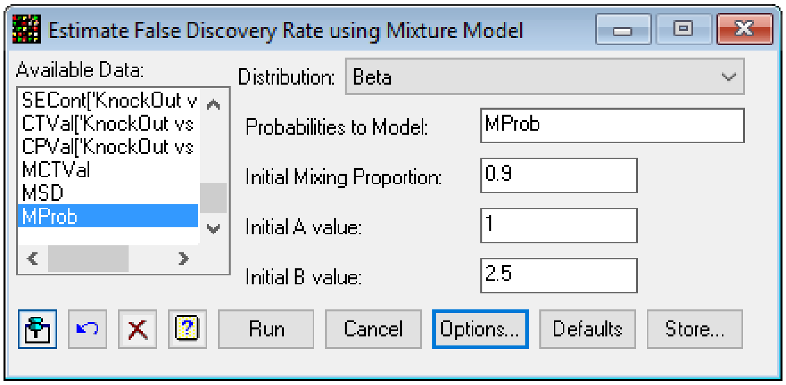
Row	Probes	Est['Reference']	Est['KnockOut']	Est['Normal']	DF	Res SD
1	Apo AI, lipid-Img	0.821786	-1.92625	1.10447	14	0.316207
2	EST, Highly similar to APOLIPOPROTEIN A-	0.817156	-1.88217	1.06501	14	0.483909
3	CATECHOL O-METHYLTRANSFERASE, MEMBRANE-B	0.696765	-1.28183	0.585061	14	0.305135
4	EST, Weakly similar to C-5 STEROL DESATU	0.431118	-0.70505	0.273932	14	0.152344
5	ESTs, Highly similar to APOLIPOPROTEIN C	0.313435	-0.620596	0.307161	14	0.172365
6	similar to yeast sterol desaturase, lipi	0.612378	-0.733336	0.120958	14	0.266376
7	Apo CIII, lipid-Img	0.397725	-0.608696	0.210971	14	0.187753

To adjust the estimated standard errors to each gene using the information across all genes the Empirical Bayes Error Estimation menu can be used. This shrinks the standard errors towards an estimated prior distribution, making the t values and probabilities more stable. To open this menu, select Empirical Bayes Error Estimation from the Analyse section of the Microarrays menu, and complete the dialog as shown to the right.



Data Type: t-values
t-values: CTVal[1]
Standard Deviations: SECont[1]
Degrees of Freedom: DF
Modified t-values: MCTVal
Modified SDs: MSD
Modified Probabilities: MProb
☑ Display in Spreadsheet

The false discovery rate can be examined by selecting False Discovery Rate from the Analyse item on the Microarrays menu. The menu right shows the settings that can be used to obtain this, and the resulting graphs are shown below


Estimate False Discovery Rate using Mixture Model

Available Data:
SECont['KnockOut v
CTVal['KnockOut vs
CPVal['KnockOut vs
MCTVal
MSD
MProb

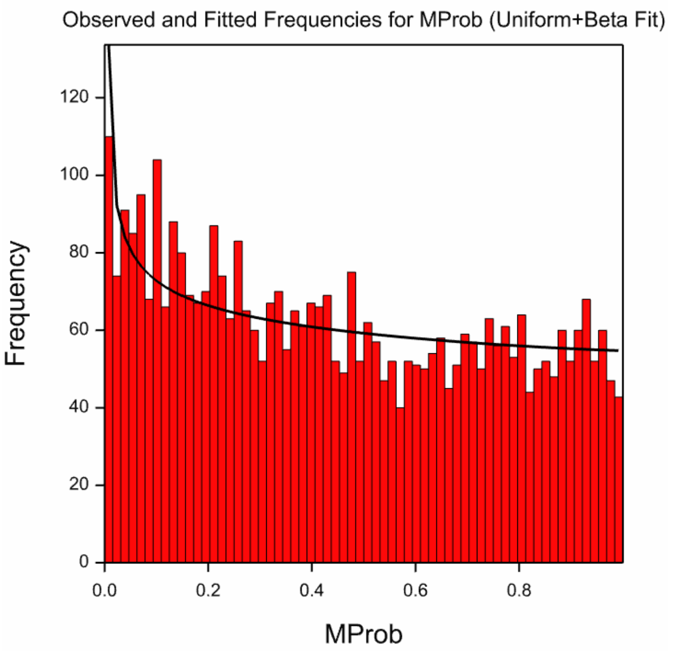
Distribution: Beta

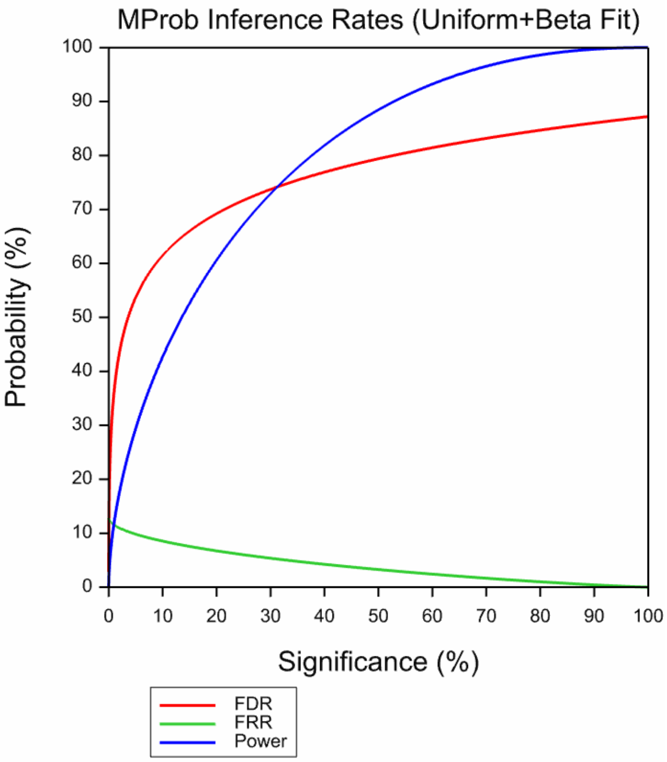
Probabilities to Model: MProb

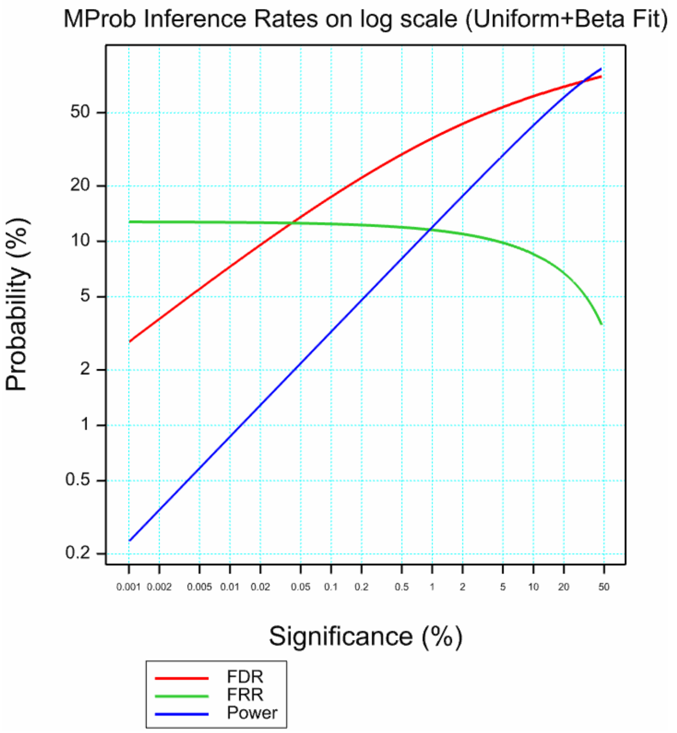
Initial Mixing Proportion: 0.9

Initial A value: 1

Initial B value: 2.5



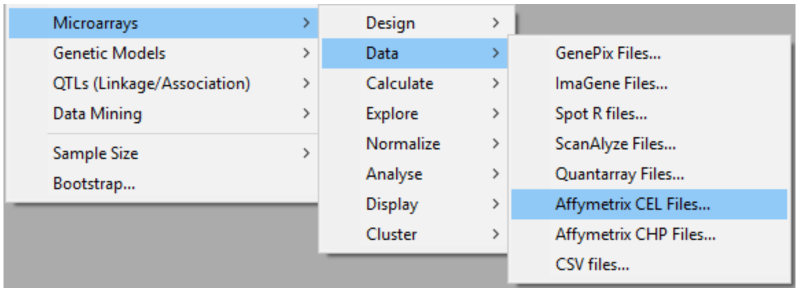




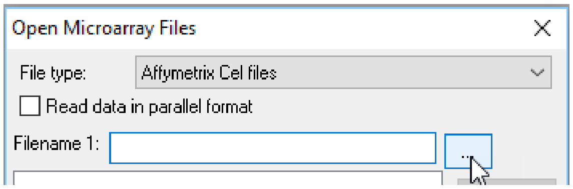
Affymetrix microarray example

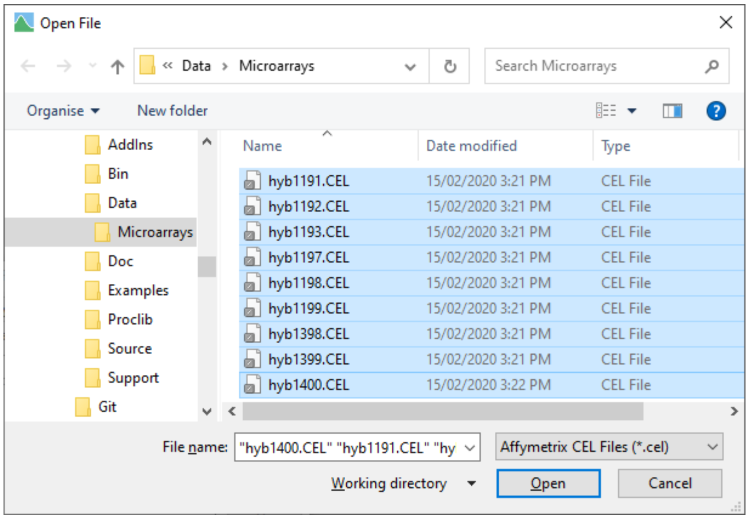
Arabidopis is a simple plant often used in gene studies. Affymetrix Arabidopis chips (ATH1-121501) have 22810 probe sets arranged in a 712 x 712 grid. In this experiment, nine of these chips (slides) were used. The CEL file data for these chips are stored in the files “hyb1191.CEL”- “hyb11400.CEL” that can be found in the Microarrays folder (see page 2). The layout of probes and quality control units (not used in the analysis) can be found in the CDF file “ATH1-121501B.CDF”. The nine slides have three replicates of three targets applied to them.

To calculate expression values for these slides, open them with the Stats | Microarrays | Data | Affymetrix CEL files menu item as shown top the right.



Now select the CEL files by using the Browse button ... as shown below.



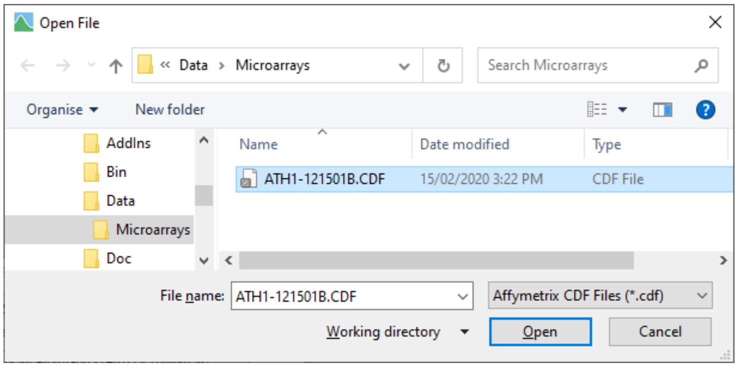


Name	Date modified	Type
hyb1191.CEL	15/02/2020 3:21 PM	CEL File
hyb1192.CEL	15/02/2020 3:21 PM	CEL File
hyb1193.CEL	15/02/2020 3:21 PM	CEL File
hyb1197.CEL	15/02/2020 3:21 PM	CEL File
hyb1198.CEL	15/02/2020 3:21 PM	CEL File
hyb1199.CEL	15/02/2020 3:21 PM	CEL File
hyb1398.CEL	15/02/2020 3:21 PM	CEL File
hyb1399.CEL	15/02/2020 3:21 PM	CEL File
hyb1400.CEL	15/02/2020 3:22 PM	CEL File

File name: "hyb1400.CEL" "hyb1191.CEL" "hy
Affymetrix CEL Files (*.cel)

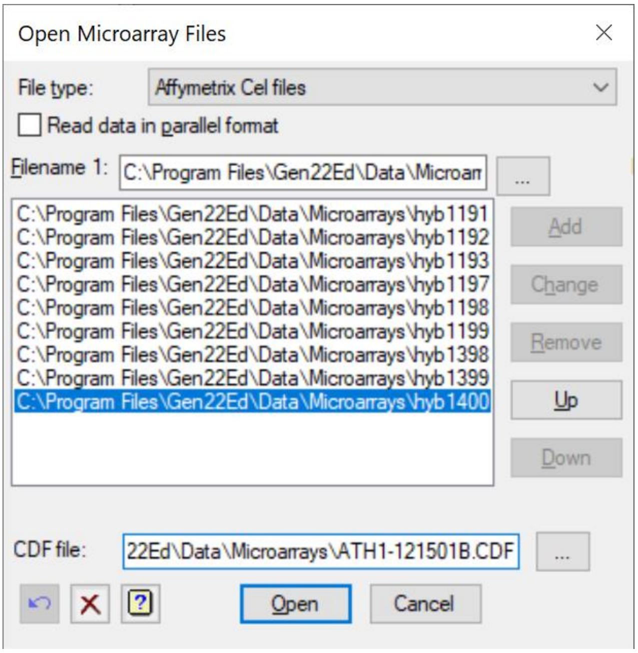
Select the nine CEL files as shown to the right.

Click the Browse button by the CDF file entry ... to select the CDF file as shown below.



Name	Date modified	Type
ATH1-121501B.CDF	15/02/2020 3:22 PM	CDF File

File name: ATH1-121501B.CDF
Affymetrix CDF Files (*.cdf)



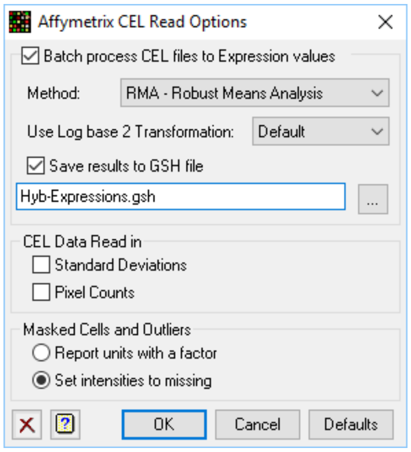
File type: Affymetrix Cel files
□ Read data in parallel format
Filename 1: C:\Program Files\Gen22Ed\Data\Microarr

C:\Program Files\Gen22Ed\Data\Microarrays\hyb1191
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1192
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1193
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1197
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1198
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1199
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1398
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1399
C:\Program Files\Gen22Ed\Data\Microarrays\hyb1400

CDF file: 22Ed\Data\Microarrays\ATH1-121501B.CDF

You may need to use the **Up** and **Down** buttons to move selected CEL files to the correct location in the list. The order the files end up in the list depends on how they are selected in the open dialog, but Windows does not always seem to give a logical ordering to these. This should give you the completed menu to the right.

Click the **Open** button and the dialog to the right will appear.



Affymetrix CEL Read Options

☑ Batch process CEL files to Expression values

Method: RMA - Robust Means Analysis

Use Log base 2 Transformation: Default

☑ Save results to GSH file

Hyb-Expressions.gsh

CEL Data Read in

☐ Standard Deviations

☐ Pixel Counts

Masked Cells and Outliers

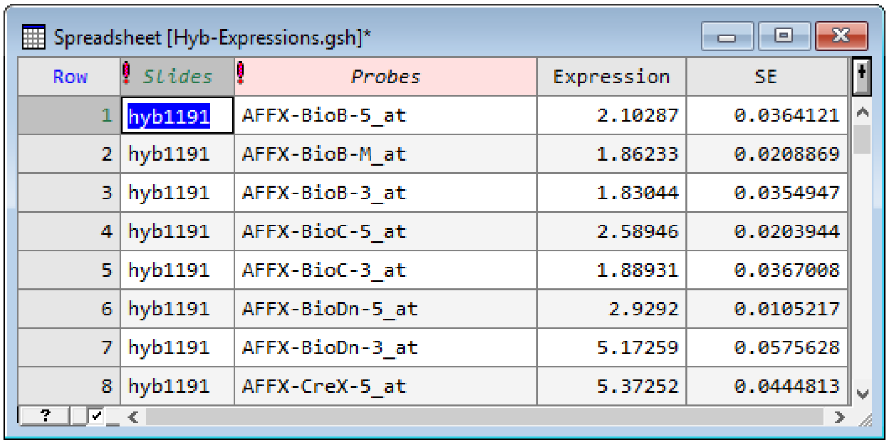
○ Report units with a factor

● Set intensities to missing

Select the option for batch processing with the RMA method and provide a filename to save the results. If no file name is provided, the results will be popped up in a spreadsheet. When you click **OK**, Genstat will produce the spreadsheet specified. This analysis can be very slow, as each CEL file contains over half a million observations.

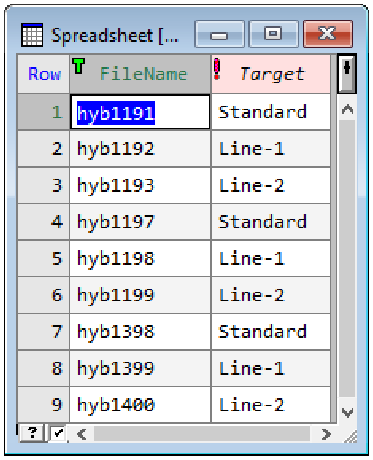
If you want to skip these steps, you can just open the already saved spreadsheet, “Hyb-Expressions.gsh” in the Microarrays folder.

Opening “Hyb-Expression.gsh” should give you the following spreadsheet:



Spreadsheet [Hyb-Expressions.gsh]*

Row	Slides	Probes	Expression	SE
1	hyb1191	AFFX-BioB-5_at	2.10287	0.0364121
2	hyb1191	AFFX-BioB-M_at	1.86233	0.0208869
3	hyb1191	AFFX-BioB-3_at	1.83044	0.0354947
4	hyb1191	AFFX-BioC-5_at	2.58946	0.0203944
5	hyb1191	AFFX-BioC-3_at	1.88931	0.0367008
6	hyb1191	AFFX-BioDn-5_at	2.9292	0.0105217
7	hyb1191	AFFX-BioDn-3_at	5.17259	0.0575628
8	hyb1191	AFFX-CreX-5_at	5.37252	0.0444813

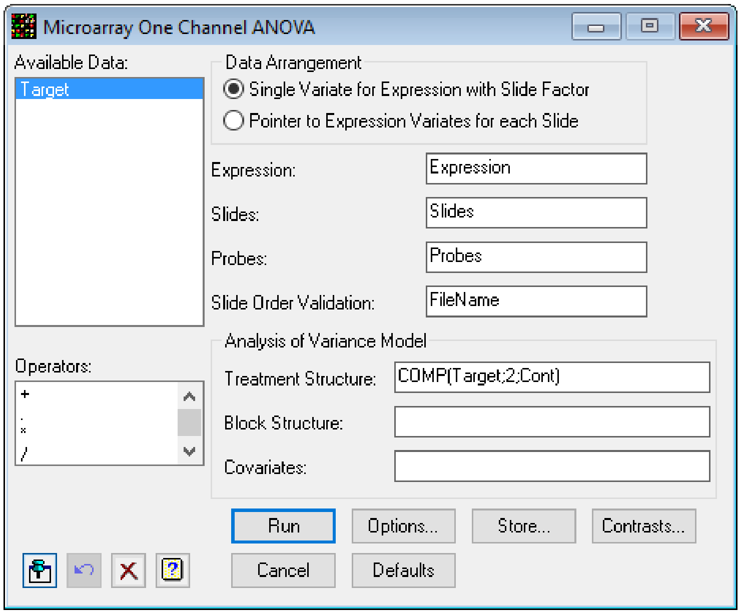


Spreadsheet [...

Row	FileName	Target
1	hyb1191	Standard
2	hyb1192	Line-1
3	hyb1193	Line-2
4	hyb1197	Standard
5	hyb1198	Line-1
6	hyb1199	Line-2
7	hyb1398	Standard
8	hyb1399	Line-1
9	hyb1400	Line-2

This data may now be summarized with the Stats | Microarrays | Analyse | One Channel ANOVA menu item, but first, we need the structure of the Targets applied to the slides. This is found in the file “HybFiles.gsh”. Open this with the File | Open menu. You should get the sheet shown to the above right.

Now open the One Channel ANOVA menu and fill in the details as shown below.



Microarray One Channel ANOVA

Available Data: Target

Data Arrangement

● Single Variate for Expression with Slide Factor

○ Pointer to Expression Variates for each Slide

Expression: Expression

Slides: Slides

Probes: Probes

Slide Order Validation: FileName

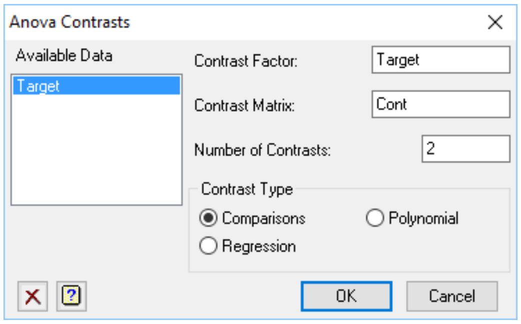
Analysis of Variance Model

Treatment Structure: COMP(Target;2;Cont)

Block Structure:

Covariates:

To estimate the difference between the Standard treatment and the other two cell lines, we can specify a contrast, using the Contrast button. This opens the menu to the right.



Anova Contrasts

Available Data: Target

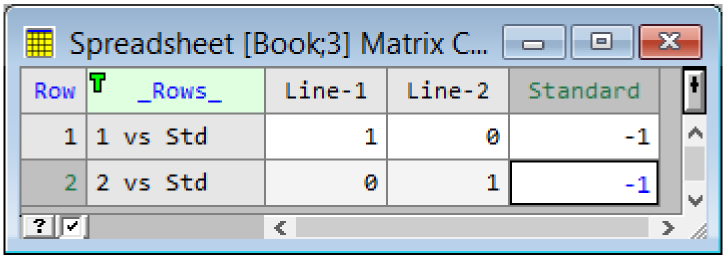
Contrast Factor: Target

Contrast Matrix: Cont

Number of Contrasts: 2

Contrast Type: ● Comparisons ○ Polynomial ○ Regression

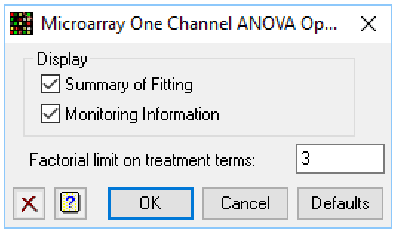
Completing the menu as shown right and clicking **OK** creates a spreadsheet containing a contrast matrix. Fill this matrix in as shown below by clicking in the various cells are typing in the entries.



Spreadsheet [Book;3] Matrix C...

Row	_Rows_	Line-1	Line-2	Standard
1	1 vs Std	1	0	-1
2	2 vs Std	0	1	-1

Now go back to the Single Channel ANOVA menu (either by clicking on it if you can see it or else by selecting it from the Windows menu) and set the options and results to be saved. Click the **Options** button and complete the dialog (right).



Microarray One Channel ANOVA Op...

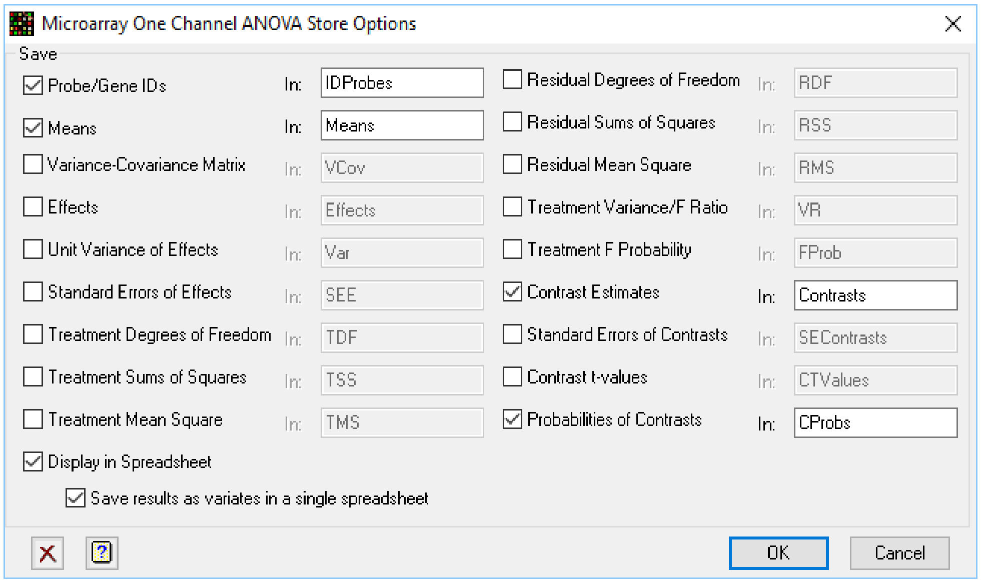
Display

☑ Summary of Fitting

☑ Monitoring Information

Factorial limit on treatment terms: 3

To save the results in a spreadsheet, click the **Store** button and complete the Store dialog as shown below.



Microarray One Channel ANOVA Store Options

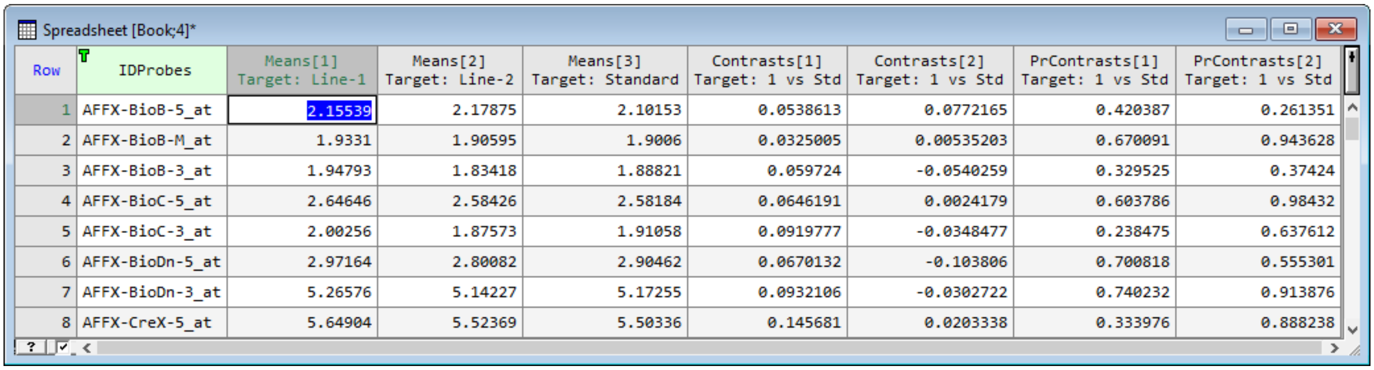
Save

☑ Probe/Gene IDs	In: IDProbes	□ Residual Degrees of Freedom	In: RDF
☑ Means	In: Means	□ Residual Sums of Squares	In: RSS
□ Variance-Covariance Matrix	In: VCov	□ Residual Mean Square	In: RMS
□ Effects	In: Effects	□ Treatment Variance/F Ratio	In: VR
□ Unit Variance of Effects	In: Var	□ Treatment F Probability	In: FProb
□ Standard Errors of Effects	In: SEE	☑ Contrast Estimates	In: Contrasts
□ Treatment Degrees of Freedom	In: TDF	□ Standard Errors of Contrasts	In: SEContrasts
□ Treatment Sums of Squares	In: TSS	□ Contrast t-values	In: CTValues
□ Treatment Mean Square	In: TMS	☑ Probabilities of Contrasts	In: CProbs

☑ Display in Spreadsheet

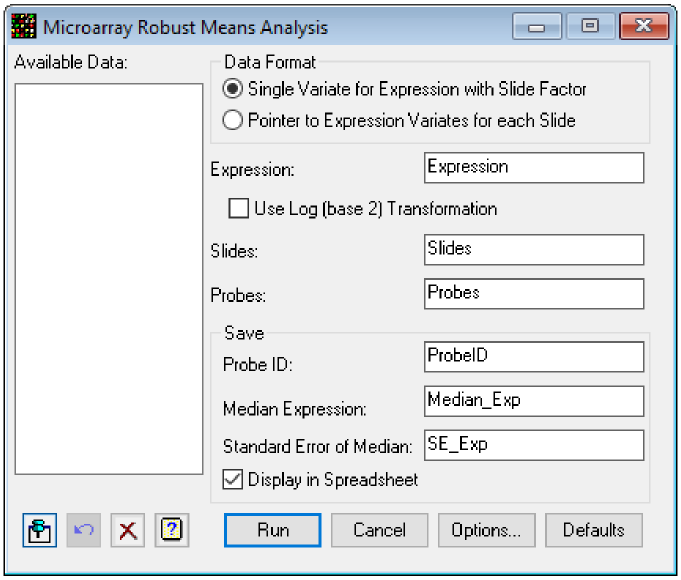
☑ Save results as variates in a single spreadsheet

When you click **Run** on the Single Channel ANOVA menu, Genstat carry out ANOVAs for each probe and pop up the results in a spreadsheet as below.


Spreadsheet [Book;4]*

Row	IDProbes	Means[1] Target: Line-1	Means[2] Target: Line-2	Means[3] Target: Standard	Contrasts[1] Target: 1 vs Std	Contrasts[2] Target: 1 vs Std	PrContrasts[1] Target: 1 vs Std	PrContrasts[2] Target: 1 vs Std
1	AFFX-BioB-5_at	2.15539	2.17875	2.10153	0.0538613	0.0772165	0.420387	0.261351
2	AFFX-BioB-M_at	1.9331	1.90595	1.9006	0.0325005	0.00535203	0.670091	0.943628
3	AFFX-BioB-3_at	1.94793	1.83418	1.88821	0.059724	-0.0540259	0.329525	0.37424
4	AFFX-BioC-5_at	2.64646	2.58426	2.58184	0.0646191	0.0024179	0.603786	0.98432
5	AFFX-BioC-3_at	2.00256	1.87573	1.91058	0.0919777	-0.0348477	0.238475	0.637612
6	AFFX-BioDn-5_at	2.97164	2.80082	2.90462	0.0670132	-0.103806	0.700818	0.555301
7	AFFX-BioDn-3_at	5.26576	5.14227	5.17255	0.0932106	-0.0302722	0.740232	0.913876
8	AFFX-CreX-5_at	5.64904	5.52369	5.50336	0.145681	0.0203338	0.333976	0.888238

Alternatively, this data can be analysed with the Robust Means Analysis menu as shown to the right, using these options shown below.


Microarray Robust Means Analysis

Available Data:

Data Format

● Single Variate for Expression with Slide Factor

○ Pointer to Expression Variates for each Slide

Expression: Expression

☐ Use Log (base 2) Transformation

Slides: Slides

Probes: Probes

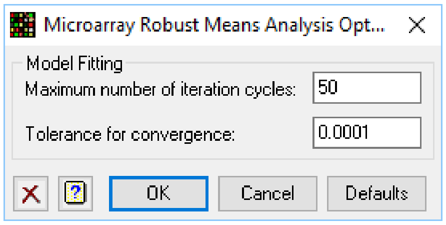
Save

Probe ID: ProbeID

Median Expression: Median_Exp

Standard Error of Median: SE_Exp

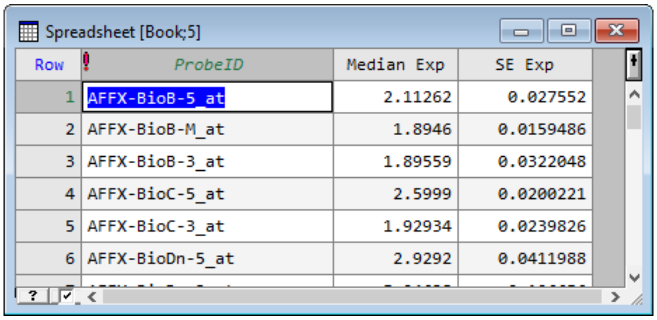
☑ Display in Spreadsheet


Microarray Robust Means Analysis Opt...

Model Fitting

Maximum number of iteration cycles: 50

Tolerance for convergence: 0.0001


Spreadsheet [Book;5]

Row	ProbeID	Median Exp	SE Exp
1	AFFX-BioB-5_at	2.11262	0.027552
2	AFFX-BioB-M_at	1.8946	0.0159486
3	AFFX-BioB-3_at	1.89559	0.0322048
4	AFFX-BioC-5_at	2.5999	0.0200221
5	AFFX-BioC-3_at	1.92934	0.0239826
6	AFFX-BioDn-5_at	2.9292	0.0411988

References

Allison, D.B., Gadbury, G.L., Heo, M., Fernandez, J.R., Lee, C.-K., Prolla, T.A., & Weindruch R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, **39**, 1-16.

Baird, D.B., Johnstone, P., and Wilson T. (2004). Normalization of microarray data using a spatial mixed model analysis which includes splines. *BioInformatics*, **20**, 3196–3205.

Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*, **64**, 479-498.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Pend, V., Ngai, J., and Speed, T.P. (2002) Normalization of cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15:1–e15:11.