Robin Users' Manual

October 2009

Table of Contents

1 Introduction

Robin represents an easy to use graphical interface for microarray (Affymetrix GeneChip, other single channel (e.g. Agilent) and two color) analysis functions from R/BioConductor. It is available on all Java-enabled computer platforms that are also supported by the R Development team. The main objective of Robin is enabling the individual biologist to use state of the art microarray preprocessing and analysis tools that are provided by the BioConductor project without in-depth knowledge of programming in R. To this end Robin provides documented, standard workflows for the quality assessment, normalization and statistical analysis of microarray data originating from many commonly used technical platforms. These workflows should allow for the analysis of most experimental setups that are conducted in microarray experiments carried out in labs around the world.

This manual gives a detailed guideline through the Robin analysis workflows for different types of microarray experiments (e.g. Affymetrix, two color, Agilent single channel…) and explains the concepts and methods of quality assessment, normalization and analysis of differential gene expression. The output that is generated by Robin can directly be imported into meta-analysis tools like MapMan and PageMan for further visualization and analysis of the data in a biological context or into Microsoft Excel.

1.1 In brief: What can Robin do for you?

You can use Robin for…

(i) quality assessment of your data
(ii) normalization of your microarray data
(iii) detection of differentially expressed genes
(iv) preparation of the data for an import into MapMan and/or excel
(v) generation of informative plots on your experiment

2 Preconditions and Glossary

2.1 Commonly used Terms

Robin helps in evaluating microarrays using advanced normalization strategies and statistics from R/BioConductor. Nevertheless, please bear in mind that most statistics and most normalization techniques make some strong assumptions and have some general terminology.

When dealing with microarrays, almost always one will deal with values which have been transformed by taking the logarithm to the base of 2. The reason is, that by logging the data, the data becomes roughly normally distributed (Gauss shaped), which then allows using tests, like student's t-test, making assumption about standard deviations etc.. Unlogged data is almost always **NOT** normally distributed, meaning t-tests are **NOT** applicable (even though they might still perform reasonably well). Thus, a difference of 1 unit means a two-fold increase or decrease in expression.

Often data is not represented as treatment value and control value, but instead of **M** and **A**. Here, **M** stands for treatment minus control (on the log scale, being a division on the normal scale), and **A** stands for (treatment plus control)/2. So **M** is a measure of your treatment effect and **A** of the expression level of that gene.
(Actually another reason for using M and A values is that it is easier to see if values deviate from the zero line as if they were deviating from a line with a slope of one. Please see Figure 1)

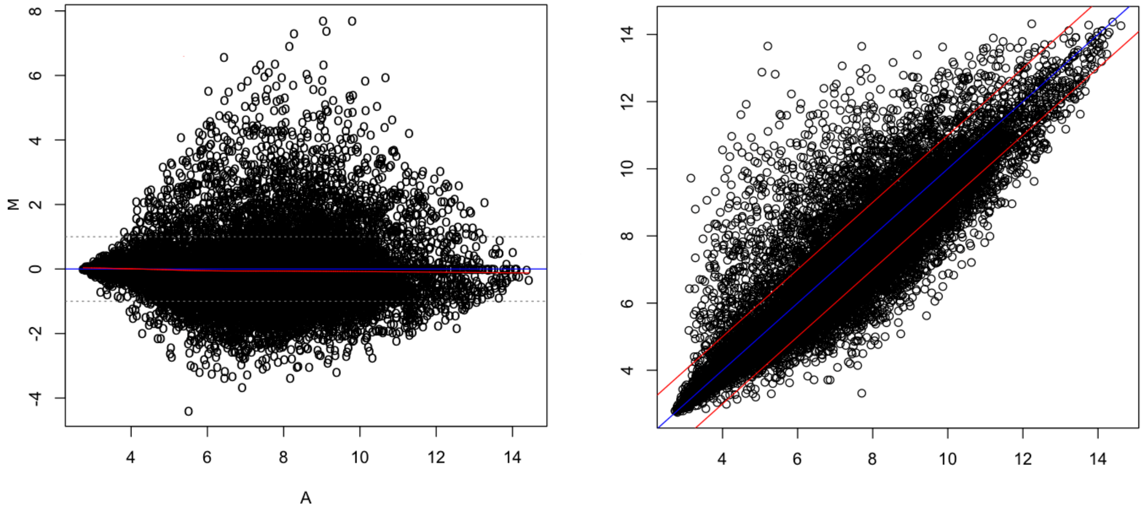


Figure 1: Comparison of MA plot versus Scatter plot of normalized expression values. The left panel shows an MA plot of the log2-fold changes when comparing two chips (M) plotted on the Y axis and the average log2 intensities (A) plotted on the X axis. On the right panel the same two chips' expression values have been plotted against each other. The MA plots gives a clearer representation of the cganges in gene expression when comparing the two chips.

2.2 Affymetrix Files

When dealing with Affymetrix chips, you will be confronted with .CEL and .CDF files, the former describes the scanned intensity for every spot (usually there are 2 times ~11 spots per gene). The CDF file describes where the spots for a probe-set are to be found on the chip, since these are not clustered to compensate for local effects such as bubbles, smears, etc.

2.3 Other single channel and two color data files

Data derived from other microarray experiments may come in a variety of different file formats depending on the microarray scanner hardware and software used. Robin supports direct import of generic file types that contain the data in text files with each column of data points separated by a special character (e.g. semicolon, TAB etc.). Import of generic data is managed by a generic data import dialog that allows you to specify which column contains what kind of data. Using this dialog it should be possible to import arbitrary microarray data. Since the generic import mechanism does not work for some data formats (like the tab-separated raw text files produced by Agilent scanners), customized settings have been supplied to allow import of these formats. Please set the import type on the file import panel according to your microarray data type if it is listed. If not, try generic import settings. If these fail we will be happy to create a customized filter for your data, if you supply us with a sample of the format. When working with generic data you'll also have to know the layout of the chips you want to analyse – presets for commonly used chip types are already included in Robin. This list can be completed with your custom layouts.

2.4 Assumptions

The strongest assumption probably being, that not much changes in your experiment. I.e. the assumption is that let's say not more than 5, 10 % of your genes are changing and that thus everything is comparable.

If this assumption is violated, you may not get satisfactory results, or worse wrong results. To demonstrate this issue, just consider the probably oldest, easiest normalization, namely median centering. Here, one just subtracts the median of one experiment from each data point. In this extreme example, Gene1 and Gene2 are completely switched off.

	XP1	XP2
Gene1	10.2	0
Gene2	3.2	0
Gene3	4.5	4.7
Gene4	7.8	7.9
Gene5	9.9	9.8
Gene6	10	10.2
median	8.85	6.3

Table 1: Experiment before normalization

	XP1	XP2
Gene1	1.35	-6.3
Gene2	-5.65	-6.3
Gene3	-4.35	-1.6
Gene4	-1.05	1.6
Gene5	1.05	3.5
Gene6	1.15	3.9

Table 2: Experiment after normalization

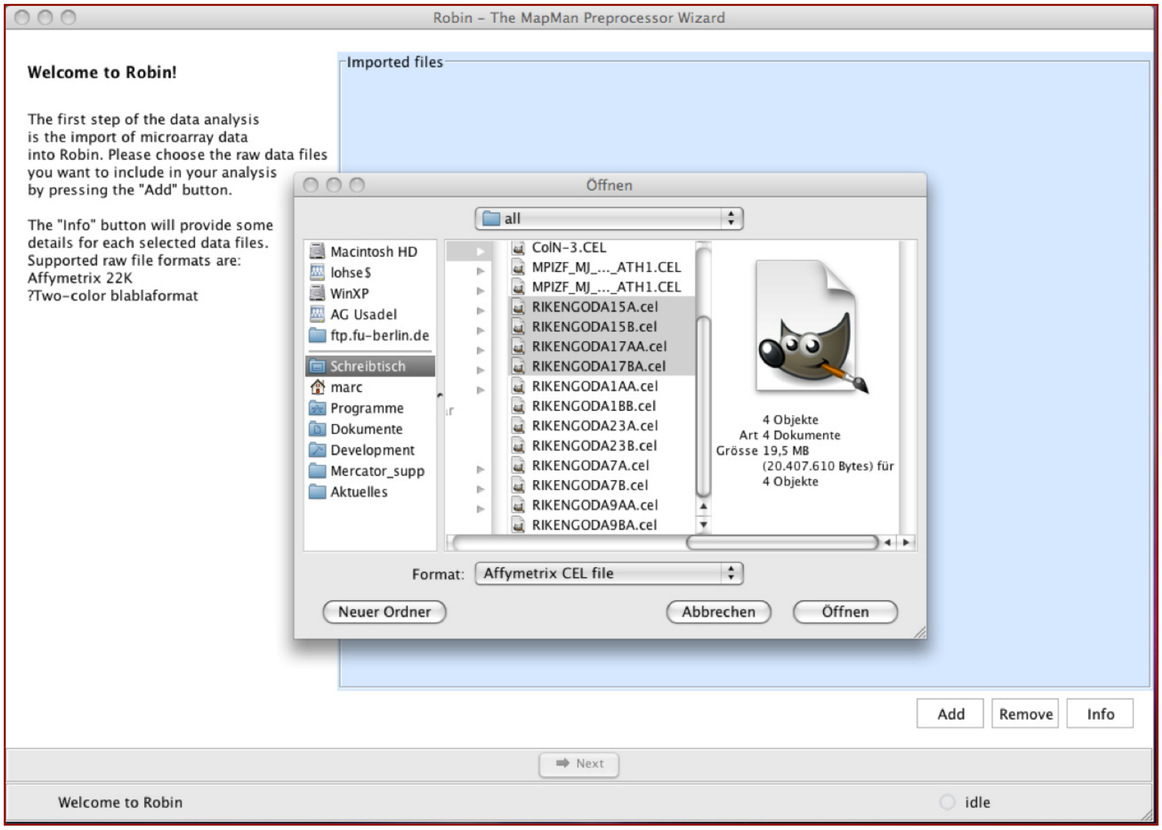
As an effect, Genes 5 and 6 seem to be upregulated, even though they were unchanged. These effects would disappear in this case, if also some genes were turned on, which often might be the case, but if you have strong suspicions, that very many genes change, and/or that these change in one direction only, you might have to consult an expert statistician.

3 Walkthroughs

The following sections of the manual provide step-by-step walkthroughs through microarray data analysis using Robin. Since Affymetrix data analysis is the most common task it is described in all detail. The workflows for two color and generic single channel analysis resemble the Affymetrix workflow and are hence described in an abbreviated fashion, focusing on the steps that are different from the Affymetrix analysis procedure.

3.1 Using Robin to analyze Affymetrix microarray data

Firstly, when using Robin, you have to localize your CEL files. Robin comes preinstalled with specialized CDF files for a small selection of organisms (arabidopsis, maize, lotus, yeast etc.), when dealing with other organisms, you will need an internet connection, so Robin can use the Bioconductor framework to install missing CDF files. The INFO button can be used to display some details about the imported CEL files such as microarray type, algorithm parameters and all the technical data included in the header section of the CEL file.



Robin – The MapMan Preprocessor Wizard

Welcome to Robin!

The first step of the data analysis is the import of microarray data into Robin. Please choose the raw data files you want to include in your analysis by pressing the "Add" button.

The "Info" button will provide some details for each selected data files. Supported raw file formats are:
Affymetrix 22K
?Two-color blablaformat

Format: Affymetrix CEL file

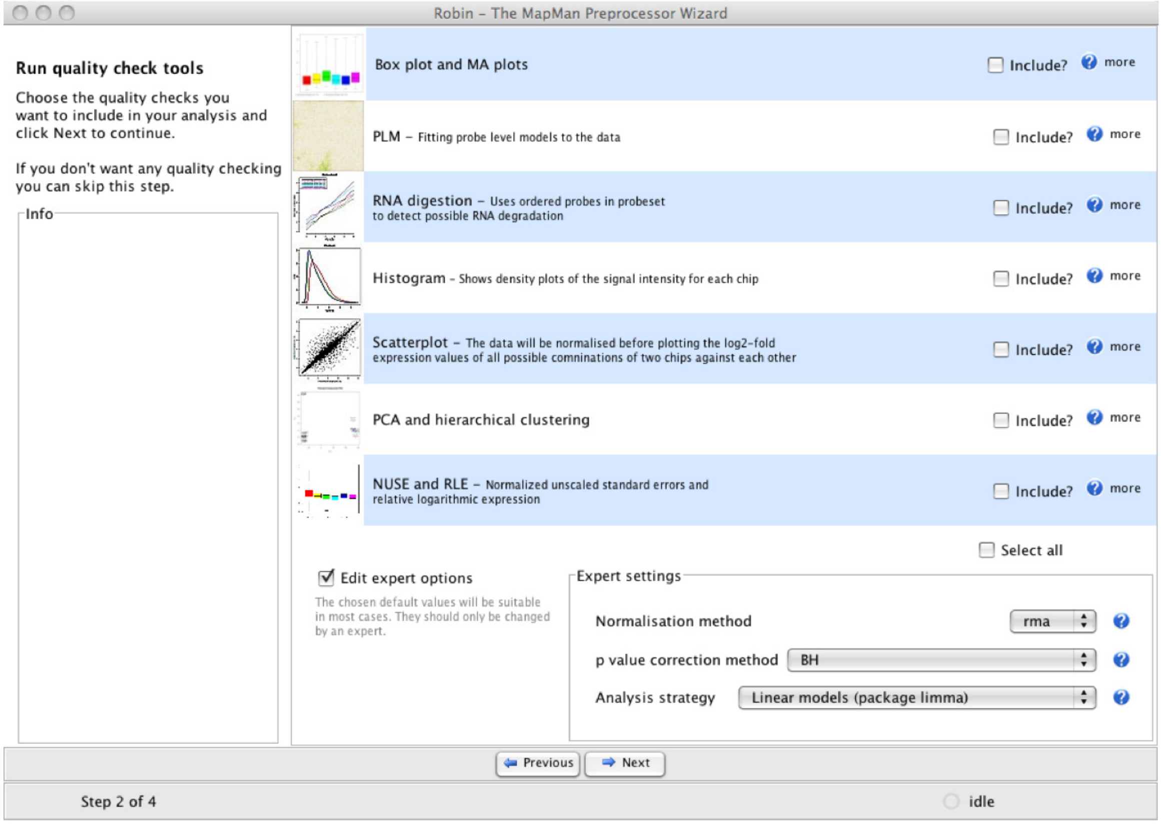
4 Objekte
Art 4 Dokumente
Grösse 19,5 MB
(20.407.610 Bytes) für
4 Objekte

Welcome to Robin

idle

Figure 2: Importing CEL files into Robin

After having selected your CEL files, you are presented with various options to investigate into the quality of the arrays.


Robin – The MapMan Preprocessor Wizard

Run quality check tools

Choose the quality checks you want to include in your analysis and click Next to continue.

If you don't want any quality checking you can skip this step.

Info

Box plot and MA plots □ Include?

PLM – Fitting probe level models to the data □ Include?

RNA digestion – Uses ordered probes in probeset to detect possible RNA degradation □ Include?

Histogram – Shows density plots of the signal intensity for each chip □ Include?

Scatterplot – The data will be normalised before plotting the log2-fold expression values of all possible comninations of two chips against each other □ Include?

PCA and hierarchical clustering □ Include?

NUSE and RLE – Normalized unscaled standard errors and relative logarithmic expression □ Include?

□ Select all

☑ Edit expert options

The chosen default values will be suitable in most cases. They should only be changed by an expert.

Expert settings

Normalisation method: rma

p value correction method: BH

Analysis strategy: Linear models (package limma)

Step 2 of 4

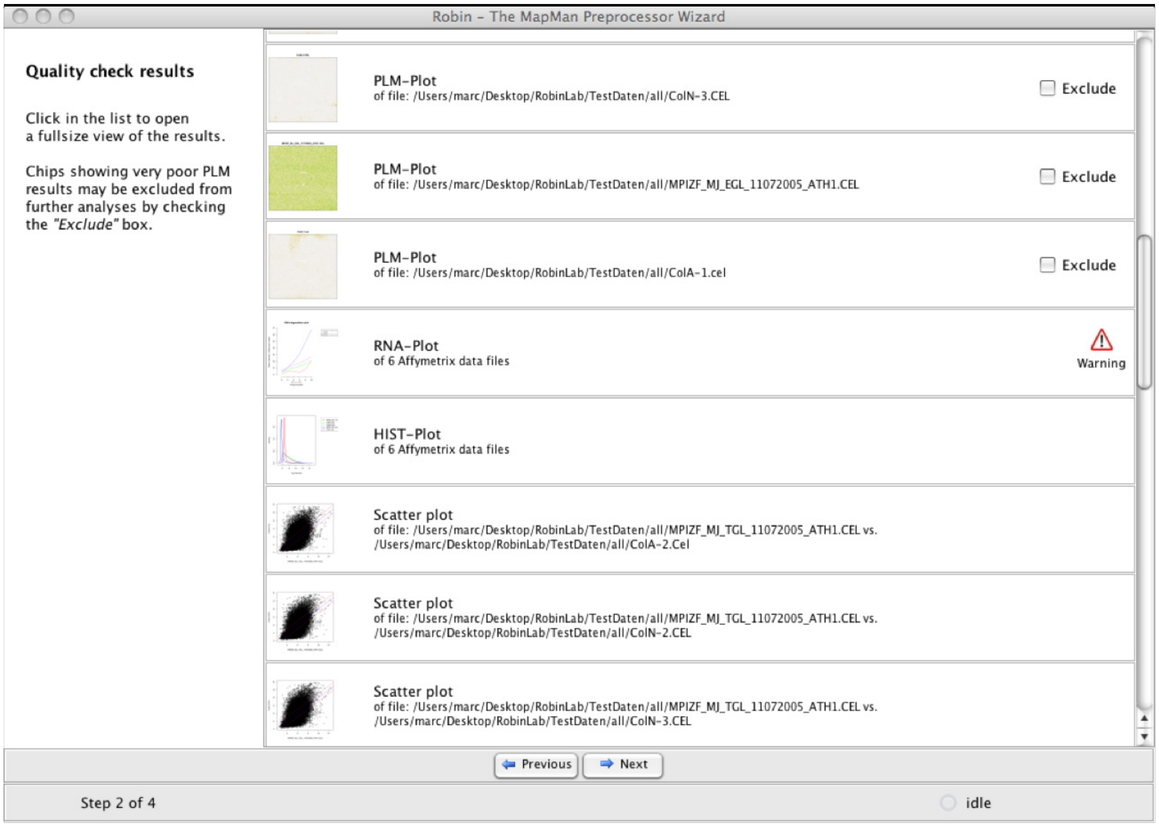
○ idle

Figure 3: Quality control options available for Affymetrix(r) arrays in Robin.

The ”expert options“ box is not shown by default – the preselected values there can be used to correctly analyze most standard experiments. If you activate the expert settings box you can explicitly choose which normalization method, p-value correction and general analysis strategy is to be used on your data.

3.1.1 Quality Control

After running the chosen quality control methods on your data, Robin will present a summary page showing thumbnails of the generated plots (see Figure 4). Clicking on the individual rows will open the images in full size and offer a possibility to save the image. **PLEASE NOTE**: You don’t have to open each image individually and save them manually – all generated quality control plots will automatically be saved together with the results of your analysis.



Robin – The MapMan Preprocessor Wizard

Quality check results

Click in the list to open
a fullsize view of the results.

Chips showing very poor PLM
results may be excluded from
further analyses by checking
the *"Exclude"* box.

PLM–Plot
of file: /Users/marc/Desktop/RobinLab/TestDaten/all/ColN-3.CEL
☐ Exclude

PLM–Plot
of file: /Users/marc/Desktop/RobinLab/TestDaten/all/MPIZF_MJ_EGL_11072005_ATH1.CEL
☐ Exclude

PLM–Plot
of file: /Users/marc/Desktop/RobinLab/TestDaten/all/ColA-1.cel
☐ Exclude

RNA–Plot
of 6 Affymetrix data files
⚠ Warning

HIST–Plot
of 6 Affymetrix data files

Scatter plot
of file: /Users/marc/Desktop/RobinLab/TestDaten/all/MPIZF_MJ_TGL_11072005_ATH1.CEL vs.
/Users/marc/Desktop/RobinLab/TestDaten/all/ColA-2.Cel

Scatter plot
of file: /Users/marc/Desktop/RobinLab/TestDaten/all/MPIZF_MJ_TGL_11072005_ATH1.CEL vs.
/Users/marc/Desktop/RobinLab/TestDaten/all/ColN-2.CEL

Scatter plot
of file: /Users/marc/Desktop/RobinLab/TestDaten/all/MPIZF_MJ_TGL_11072005_ATH1.CEL vs.
/Users/marc/Desktop/RobinLab/TestDaten/all/ColN-3.CEL

Step 2 of 4

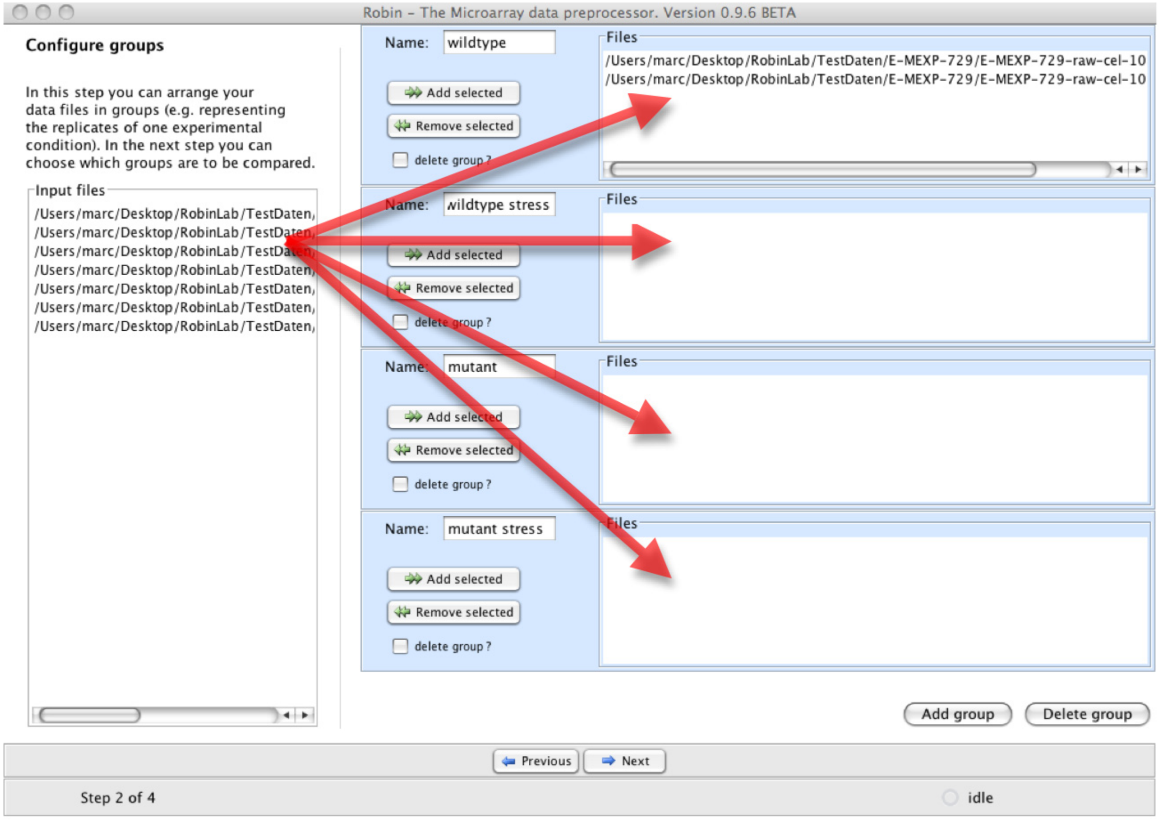
○ idle

Figure 4: Quality analysis summary page.

Some of the quality assessments functions may have issued warnings – clicking on the small warning icon will open an info panel that tells you more specifically why the warning was generated. For example the RNA degradation analysis may have identified chips that display slopes higher than the accepted threshold or whose slopes deviate by more than 10 per cent from the median slope (see section 4.1.5 for details). Individual chips displaying an extraordinarily bad quality in the PLM-Plot (see 4.1.3) or MA Plot (see 4.1.2) can be excluded from further analyses by checking the "Exclude" box. Section 4 describes all available quality control methods in detail and gives examples of good and bad quality check results.

3.1.2 Experiment design and statistical analysis

The next step in the analysis workflow is the assignment of the chips to groups of biological replicates. NOTE: Robin analyses all replicates as biological replicates – there is no way implemented yet that allows for proper consideration of technical replicates. Please be aware that if technical replicates are imported the statistical test outputs will not be sound any more. You can choose a descriptive unique name for each group of replicates (like "mutant", "wildtpye" etc : see Figure 5). After sorting the chips, clicking "next" will proceed to the graphical experiment designer. Here the user can set up the comparisons that are to be made by CTRL-click-dragging connections between the groups (see Figure 6). Direct comparisons e.g. wild type against mutant samples are defined by simply dragging an arrow from the "wildtype" to the "mutants" box on the left



Configure groups

In this step you can arrange your data files in groups (e.g. representing the replicates of one experimental condition). In the next step you can choose which groups are to be compared.

Input files

/Users/marc/Desktop/RobinLab/TestDaten/
/Users/marc/Desktop/RobinLab/TestDaten/
/Users/marc/Desktop/RobinLab/TestDaten/
/Users/marc/Desktop/RobinLab/TestDaten/
/Users/marc/Desktop/RobinLab/TestDaten/
/Users/marc/Desktop/RobinLab/TestDaten/
/Users/marc/Desktop/RobinLab/TestDaten/

Name: wildtype

Files
/Users/marc/Desktop/RobinLab/TestDaten/E-MEXP-729/E-MEXP-729-raw-cel-10
/Users/marc/Desktop/RobinLab/TestDaten/E-MEXP-729/E-MEXP-729-raw-cel-10

Name: wildtype stress

Name: mutant

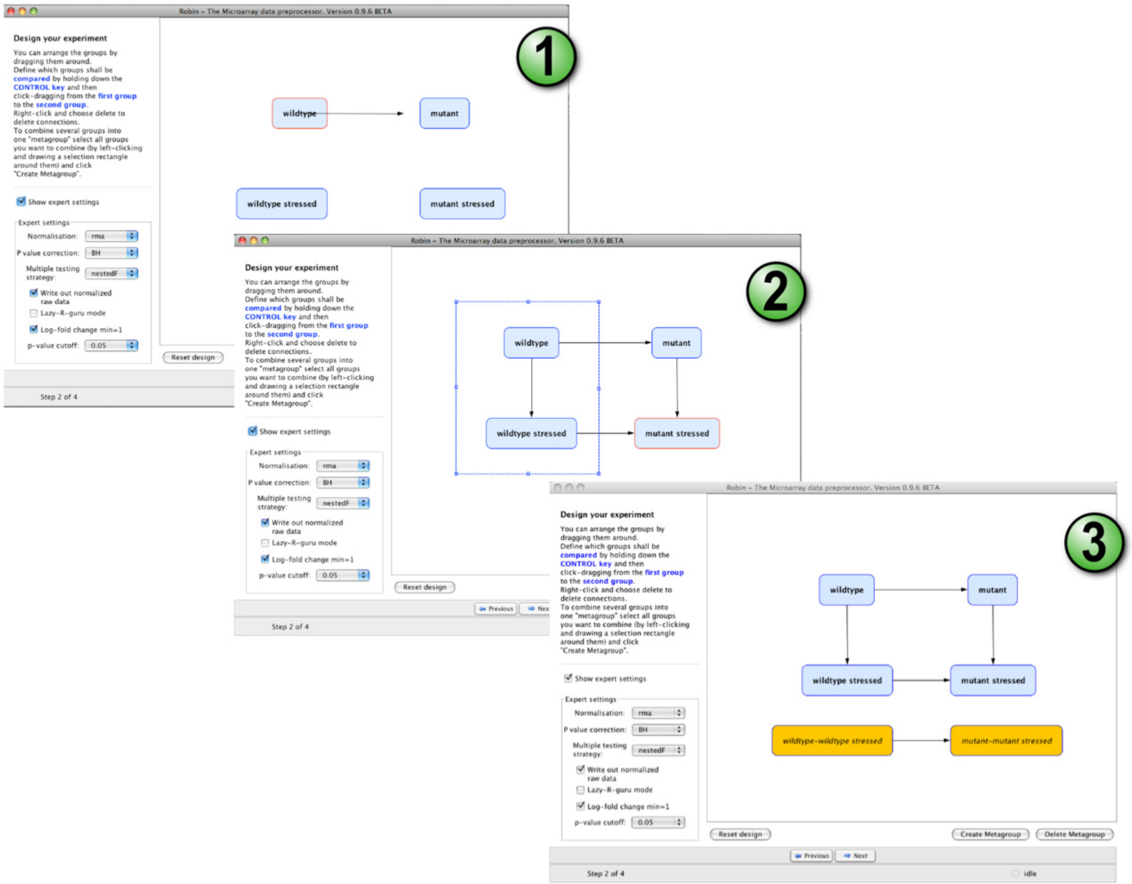
Name: mutant stress

Step 2 of 4

Figure 5: Sorting of replicate experiments into named groups.

panel of the graphical designer screen (Figure 6.1). If experiments with more than one varying experimental condition are to be analysed the user can combine groups into meta groups and define comparisons of meta groups by dragging connections between them. In the example experiment shown in Figure 6, mutant and wild type plants were compared both under stress and normal conditions - so the experiment varies in two dimensions with genotype (wild type or mutant) being one factor and treatment (stress, no stress) being the other. The first four direct comparisons (Figure 6.2) will yield the genes that are responding to the treatment in the wild type ("wildtype – wildtype stressed) and the mutant ("mutant – mutant stressed"), which genes respond differently between the genotypes under normal conditions ("wildtype – mutant") and stress and which genes generally respond differentially in the two genotypes ("(wildtype – wildtype stressed) – (mutant – mutant stressed") – this is also referred to as the *interaction term;* see Figure 6.3).

PLEASE NOTE: The direction of the arrow specifies the direction of the comparison. When the arrow points from the wildtype to the mutant this should be read as "wildtype minus mutant". Genes showing a higher expression in the mutant when compared to the wildtype will accordingly yield a negative log2-fold change value as a result!



Robin – The Microarray data preprocessor. Version 0.9.6 BETA

Design your experiment

You can arrange the groups by dragging them around.
Define which groups shall be **compared** by holding down the **CONTROL key** and then click-dragging from the **first group** to the **second group**.
Right-click and choose delete to delete connections.
To combine several groups into one "metagroup" select all groups you want to combine (by left-clicking and drawing a selection rectangle around them) and click "Create Metagroup".

☑ Show expert settings

Expert settings
Normalisation: rma
P value correction: BH
Multiple testing strategy: nestedF
☑ Write out normalized raw data
☐ Lazy-R-guru mode
☑ Log-fold change min=1
p-value cutoff: 0.05

Step 2 of 4

idle

Figure 6: Setting up the experiment using Robins graphical designer.

The experiment designer panel also offers an expert option box that enables the experienced user to influence specific parameters of the statistical inference. By default, the normalization method used for the main analysis will be the same that was chosen on the quality check panel (see Figure 6; if nothing was changed the default will be robust multi array averaging - RMA) to ensure consistency between the quality check and differential expression statistics. The user can define significance cut-offs like discarding genes that show a log2 fold change in expression lesser than 1 (i.e. less than 2-fold up- or down regulation) and genes showing a p-value greater than e.g. 0.05 (i.e. 5% false discovery rate is accepted). A choice of multiple testing methods is available for the inference of differentially expressed genes:

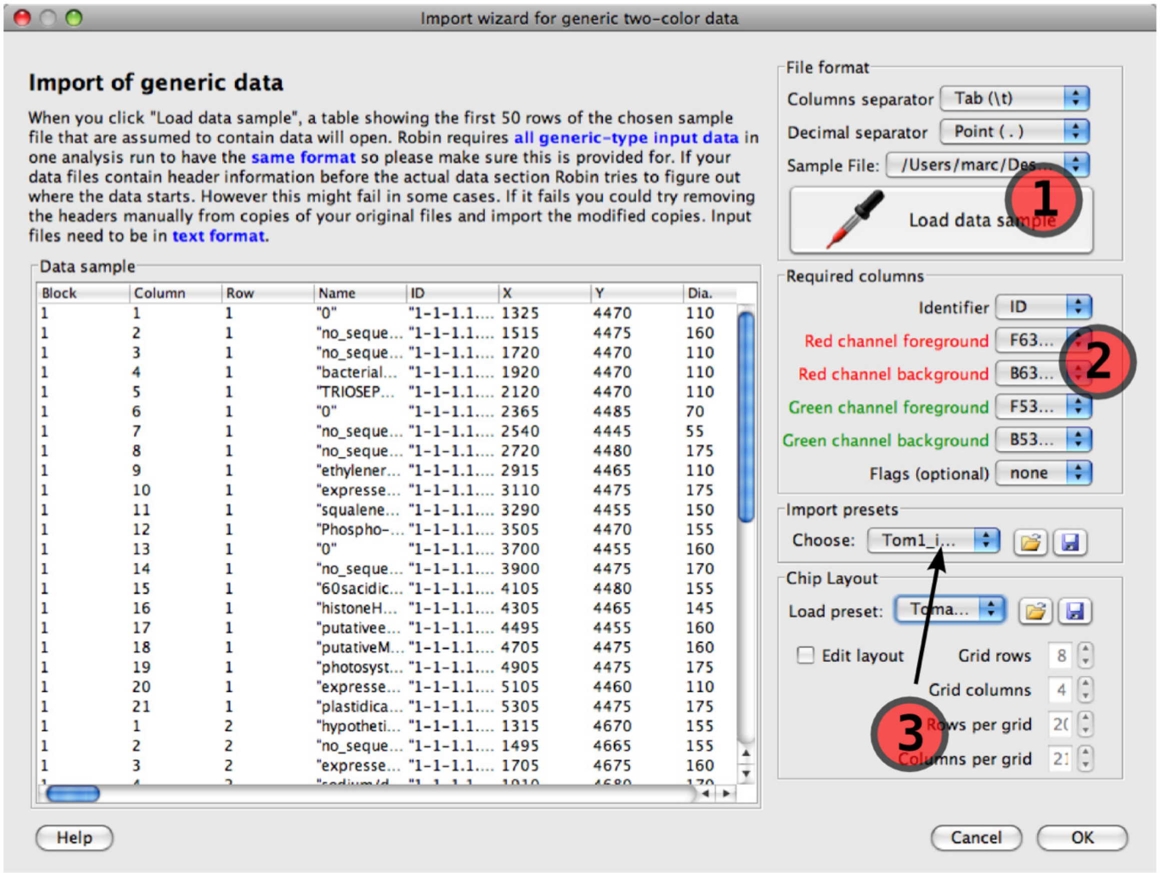
1) “separate” – Does the multiple testing for each comparison (contrast) separately. Using this method, each specific comparison will always give the same result irrespective of the set of comparisons being made in the analysis. It is the simplest method available as it does not consider multiple testing adjustment between the comparisons and assumes the same raw p-value cut-off for all comparisons (which might be very different).

2) “global” - Implements multiple testing correction across all comparisons and probes simultaneously ensuring a consistent p-value cut-off across all comparisons.

3) “hierarchical” – Does p-value adjustment first for all genes and then across comparisons, which offers more statistical power to control the family-wise error rate when using the method described by (Holm, 1979) for p-value adjustment.

4) “nestedF” – First does p-value adjustment for all genes and uses a nested F test to classify the comparisons as significant or not for the selected genes.

Users that are familiar with R programming can activate the “preview R script”-mode in which all scripts generated by Robin are shown in an internal editor for review and modification prior to execution. Even if this option was not chosen, all R scripts generated by Robin will be written to the “source” folder in the final output directory. When the design step is completed, clicking the “Next” button will first open a file browser asking for a location to save the results to and then move on to the execution of the analysis. After completing the calculations, Robin will show a summary of the warnings generated during the workflow (if any) and offer options to exit, restart or modify the current experiment.



Import wizard for generic two-color data

Import of generic data

When you click "Load data sample", a table showing the first 50 rows of the chosen sample file that are assumed to contain data will open. Robin requires **all generic-type input data** in one analysis run to have the **same format** so please make sure this is provided for. If your data files contain header information before the actual data section Robin tries to figure out where the data starts. However this might fail in some cases. If it fails you could try removing the headers manually from copies of your original files and import the modified copies. Input files need to be in **text format**.

Data sample

Block	Column	Row	Name	ID	X	Y	Dia.
1	1	1	"0"	"1-1-1.1....	1325	4470	110
1	2	1	"no_seque...	"1-1-1.1....	1515	4475	160
1	3	1	"no_seque...	"1-1-1.1....	1720	4470	110
1	4	1	"bacterial...	"1-1-1.1....	1920	4470	110
1	5	1	"TRIOSEP...	"1-1-1.1....	2120	4470	110
1	6	1	"0"	"1-1-1.1....	2365	4485	70
1	7	1	"no_seque...	"1-1-1.1....	2540	4445	55
1	8	1	"no_seque...	"1-1-1.1....	2720	4480	175
1	9	1	"ethylener...	"1-1-1.1....	2915	4465	110
1	10	1	"expresse...	"1-1-1.1....	3110	4475	175
1	11	1	"squalene...	"1-1-1.1....	3290	4455	150
1	12	1	"Phospho-...	"1-1-1.1....	3505	4470	155
1	13	1	"0"	"1-1-1.1....	3700	4455	160
1	14	1	"no_seque...	"1-1-1.1....	3900	4475	170
1	15	1	"60sacidic...	"1-1-1.1....	4105	4480	155
1	16	1	"histoneH...	"1-1-1.1....	4305	4465	145
1	17	1	"putativee...	"1-1-1.1....	4495	4455	160
1	18	1	"putativeM...	"1-1-1.1....	4705	4475	160
1	19	1	"photosyst...	"1-1-1.1....	4905	4475	175
1	20	1	"expresse...	"1-1-1.1....	5105	4460	110
1	21	1	"plastidica...	"1-1-1.1....	5305	4475	175
1	1	2	"hypotheti...	"1-1-1.1....	1315	4670	155
1	2	2	"no_seque...	"1-1-1.1....	1495	4665	155
1	3	2	"expresse...	"1-1-1.1....	1705	4675	160

File format

Columns separator: Tab (\t)

Decimal separator: Point (.)

Sample File: /Users/marc/Des...

Load data sample

Required columns

Identifier: ID

Red channel foreground: F63...

Red channel background: B63...

Green channel foreground: F53...

Green channel background: B53...

Flags (optional): none

Import presets

Choose: Tom1_i...

Chip Layout

Load preset: Toma...

☐ Edit layout

Grid rows: 8

Grid columns: 4

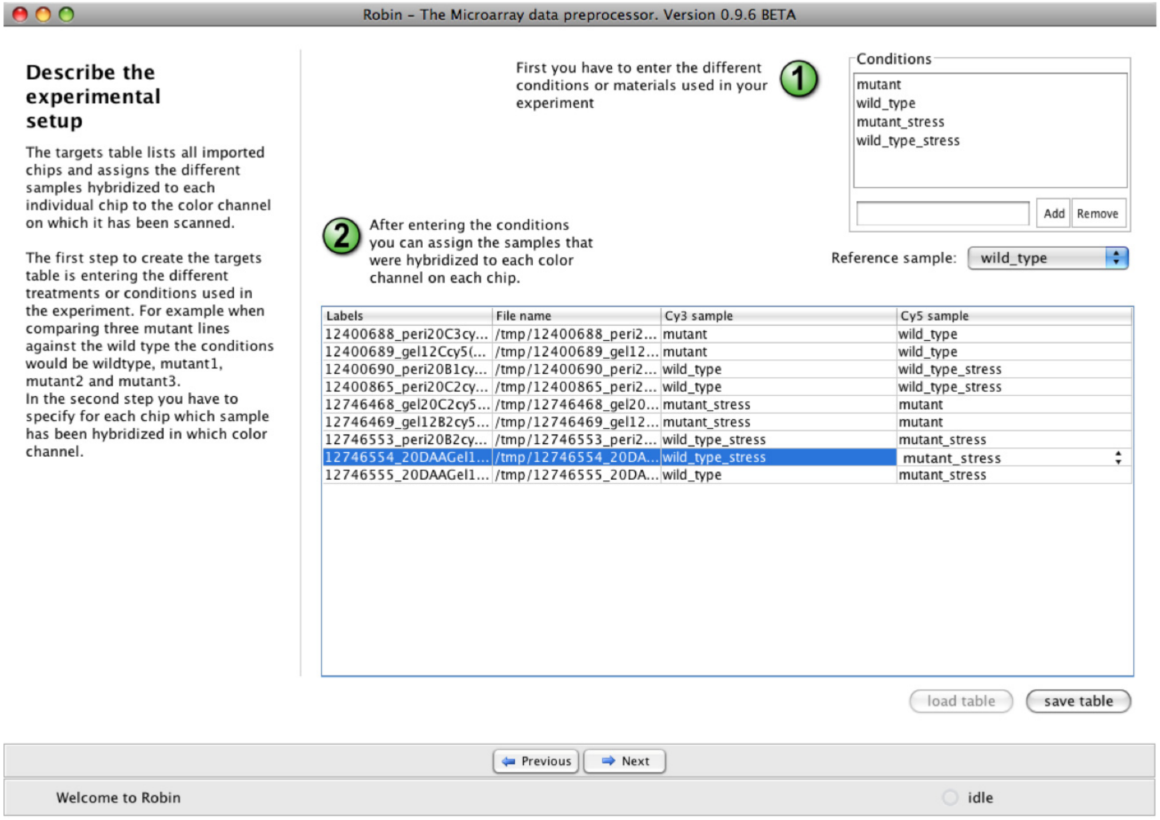
Rows per grid: 2(

Columns per grid: 21

Figure 7: Two color data import wizard. Robin automatically removes header sections from different tabular file formats and extracts the column headers. The user has to define which column contains which data (1) by assigning the proper column names to the required column fields. After choosing a chip layout from the list of layout presets (or defining a new layout and saving it as a preset; see 2), the user can save the import settings (3) and reuse them later when importing data of the same format.

3.2 Analysing two-colour microarray data

The first step when working with two colour microarray data is data import. Robin provides a wizard dialog that helps the user to import various import formats with the only restriction that the data has to be provided in plain text format (.csv, .tab etc). Loading MS Excel worksheets directly is not supported (yet). Aside from this any kind of tabular data can be used. When importing data, the user only needs to know which column separator was used. Layouts of frequently used microarray types are included as clickable presets in the layout preset list – if the layout of your favorite chips is not included in the list, you can define a new layout and save it for later use. The minimal data required to analyze two color chips is an identifier uniquely identifying the oligos / cDNAs spotted on the chip and the red and green channel foreground and background signal intensities. The table view on the left half of the import dialog facilitates choosing the column containing the required data, and after specifying the column names under "Required columns" the information needed to import the data is complete. Robin will create copies of the input files that are stripped off any header text and checked row-wise for data format consistency. The processed input files will be placed in a separate folder in the output folder.



Robin – The Microarray data preprocessor. Version 0.9.6 BETA

Describe the experimental setup

The targets table lists all imported chips and assigns the different samples hybridized to each individual chip to the color channel on which it has been scanned.

The first step to create the targets table is entering the different treatments or conditions used in the experiment. For example when comparing three mutant lines against the wild type the conditions would be wildtype, mutant1, mutant2 and mutant3.
In the second step you have to specify for each chip which sample has been hybridized in which color channel.

1 First you have to enter the different conditions or materials used in your experiment

Conditions
mutant

wild_type

mutant_stress

wild_type_stress

2 After entering the conditions you can assign the samples that were hybridized to each color channel on each chip.

Reference sample: wild_type

Labels	File name	Cy3 sample	Cy5 sample
12400688_peri20C3cy...	/tmp/12400688_peri2...	mutant	wild_type
12400689_gel12Ccy5(...	/tmp/12400689_gel12...	mutant	wild_type
12400690_peri20B1cy...	/tmp/12400690_peri2...	wild_type	wild_type_stress
12400865_peri20C2cy...	/tmp/12400865_peri2...	wild_type	wild_type_stress
12746468_gel20C2cy5...	/tmp/12746468_gel20...	mutant_stress	mutant
12746469_gel12B2cy5...	/tmp/12746469_gel12...	mutant_stress	mutant
12746553_peri20B2cy...	/tmp/12746553_peri2...	wild_type_stress	mutant_stress
12746554_20DAAGel1...	/tmp/12746554_20DA...	wild_type_stress	mutant_stress
12746555_20DAAGel1...	/tmp/12746555_20DA...	wild_type	mutant_stress

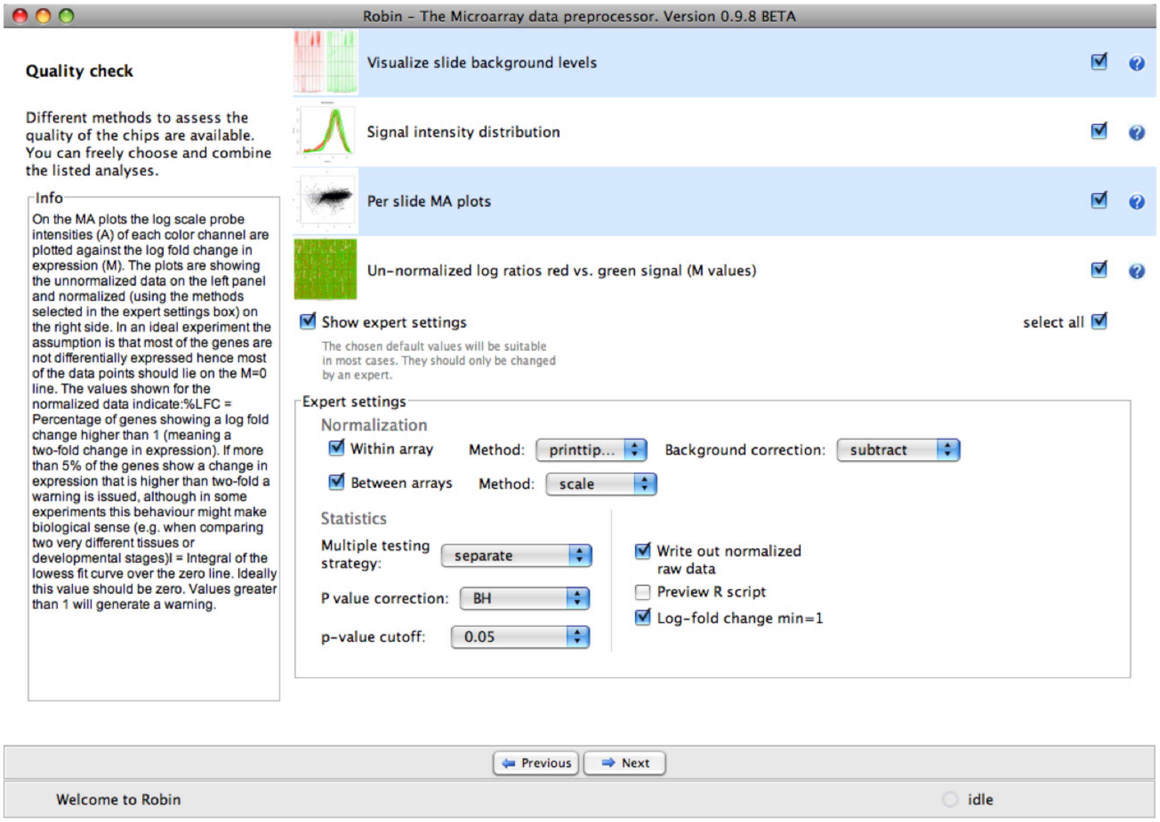
Welcome to Robin

idle

Figure 8: Defining the RNA targets table for two-color microarrays

The next piece of information Robin needs is which different RNA samples (RNA targets) have been hybridized to which channel on which chip. This can be conveniently entered on the targets table panel (see Figure 8). Robin will run some checks on the input to assert consistency. Analogous to Affymetrix data analysis the next panel provides a choice of quality check methods adapted to two color arrays and an expert settings box granting deeper control of the analysis parameters (See Figure 9).
Each step of the normalization process, namely background correction, within-array normalization and between-array normalization can be configured separately to o

Quality check results will be summarized in a list resembling Figure 4. Depending on the amount of factors being varied in the experiment (i.e. the amount of different RNA samples hybridized) clicking “Next” on the quality check panel will either directly start the main analysis (e.g. in a simple two sample comparison) or open the graphical experiment designer panel (see Figure 6). Experiment layout is done exactly as for Affymetrix arrays – please refer to section “Experiment design“ for a detailed description.



Robin – The Microarray data preprocessor. Version 0.9.8 BETA

Quality check

Different methods to assess the quality of the chips are available. You can freely choose and combine the listed analyses.

Info

On the MA plots the log scale probe intensities (A) of each color channel are plotted against the log fold change in expression (M). The plots are showing the unnormalized data on the left panel and normalized (using the methods selected in the expert settings box) on the right side. In an ideal experiment the assumption is that most of the genes are not differentially expressed hence most of the data points should lie on the M=0 line. The values shown for the normalized data indicate:%LFC = Percentage of genes showing a log fold change higher than 1 (meaning a two-fold change in expression). If more than 5% of the genes show a change in expression that is higher than two-fold a warning is issued, although in some experiments this behaviour might make biological sense (e.g. when comparing two very different tissues or developmental stages)I = Integral of the lowess fit curve over the zero line. Ideally this value should be zero. Values greater than 1 will generate a warning.

- Visualize slide background levels ☑
- Signal intensity distribution ☑
- Per slide MA plots ☑
- Un-normalized log ratios red vs. green signal (M values) ☑

☑ Show expert settings — select all ☑

The chosen default values will be suitable in most cases. They should only be changed by an expert.

Expert settings

Normalization

☑ Within array — Method: printtip... — Background correction: subtract

☑ Between arrays — Method: scale

Statistics

Multiple testing strategy: separate

P value correction: BH

p-value cutoff: 0.05

☑ Write out normalized raw data

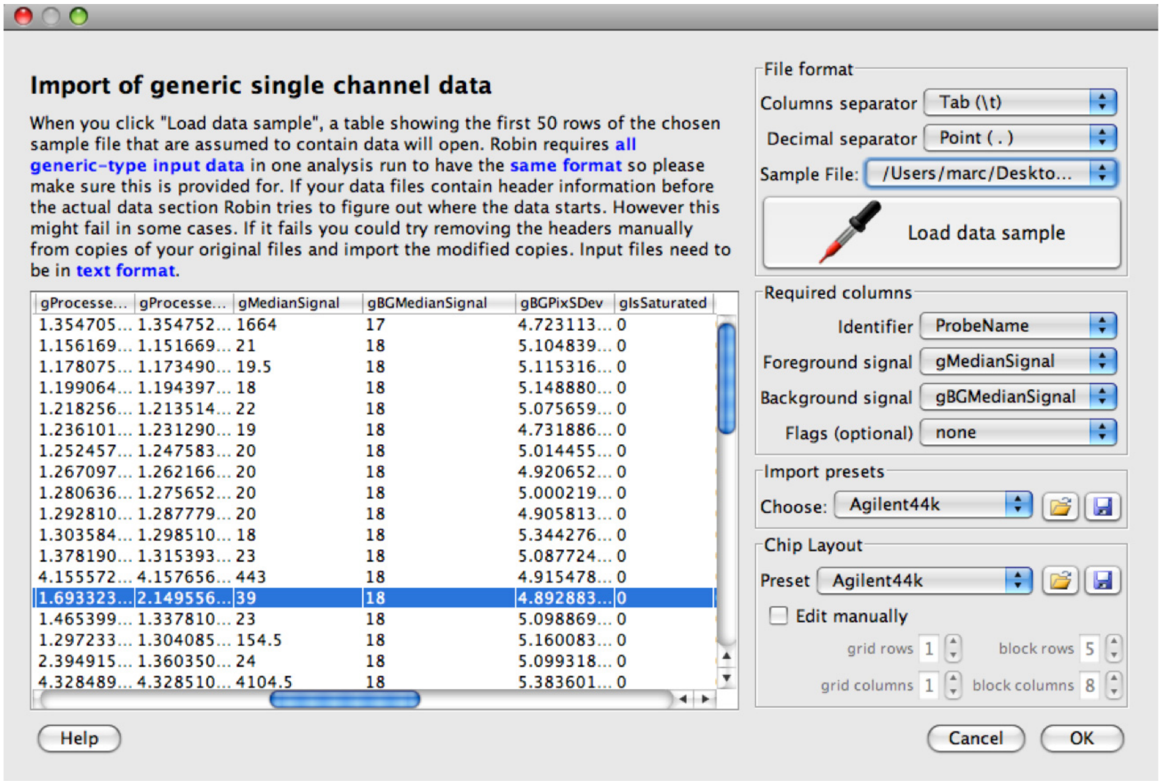
□ Preview R script

☑ Log-fold change min=1

Figure 9: Quality check and expert settings for two color microarrays

3.3 Analysis of generic single channel arrays (e.g. Agilent)

Analysis of generic single channel arrays resembles the workflow for two color chip analysis in the largest part. The flexible import dialog (see Figure 10) allows for configuration of any tabular text file based data. Please note that you have to specify whether the data originates from an Agilent scanner prior to import to make sure that Robin can correctly remove the header section of the data files. Robin will process the input according the configuration chosen in the import dialog and create cleaned-up working copies, leaving the original data untouched. In the next step, analogous to Affymetrix data analysis, several assessment methods can be chosen to investigate into the chips' quality. Since most generic single channel chips are not based on a probeset design (several probes per target) but only contain one probe per target transcript, the probeset specific quality checks available for Affymetrix arrays (i.e. PLM-based analyses, RNA degradation plot) cannot be used.



Import of generic single channel data

When you click "Load data sample", a table showing the first 50 rows of the chosen sample file that are assumed to contain data will open. Robin requires **all generic-type input data** in one analysis run to have the **same format** so please make sure this is provided for. If your data files contain header information before the actual data section Robin tries to figure out where the data starts. However this might fail in some cases. If it fails you could try removing the headers manually from copies of your original files and import the modified copies. Input files need to be in **text format**.

gProcesse...	gProcesse...	gMedianSignal	gBGMedianSignal	gBGPixSDev	gIsSaturated
1.354705...	1.354752...	1664	17	4.723113...	0
1.156169...	1.151669...	21	18	5.104839...	0
1.178075...	1.173490...	19.5	18	5.115316...	0
1.199064...	1.194397...	18	18	5.148880...	0
1.218256...	1.213514...	22	18	5.075659...	0
1.236101...	1.231290...	19	18	4.731886...	0
1.252457...	1.247583...	20	18	5.014455...	0
1.267097...	1.262166...	20	18	4.920652...	0
1.280636...	1.275652...	20	18	5.000219...	0
1.292810...	1.287779...	20	18	4.905813...	0
1.303584...	1.298510...	18	18	5.344276...	0
1.378190...	1.315393...	23	18	5.087724...	0
4.155572...	4.157656...	443	18	4.915478...	0
1.693323...	2.149556...	39	18	4.892883...	0
1.465399...	1.337810...	23	18	5.098869...	0
1.297233...	1.304085...	154.5	18	5.160083...	0
2.394915...	1.360350...	24	18	5.099318...	0
4.328489...	4.328510...	4104.5	18	5.383601...	0

File format

Columns separator: Tab (\t)

Decimal separator: Point (.)

Sample File: /Users/marc/Deskto...

Required columns

Identifier: ProbeName

Foreground signal: gMedianSignal

Background signal: gBGMedianSignal

Flags (optional): none

Import presets

Choose: Agilent44k

Chip Layout

Preset: Agilent44k

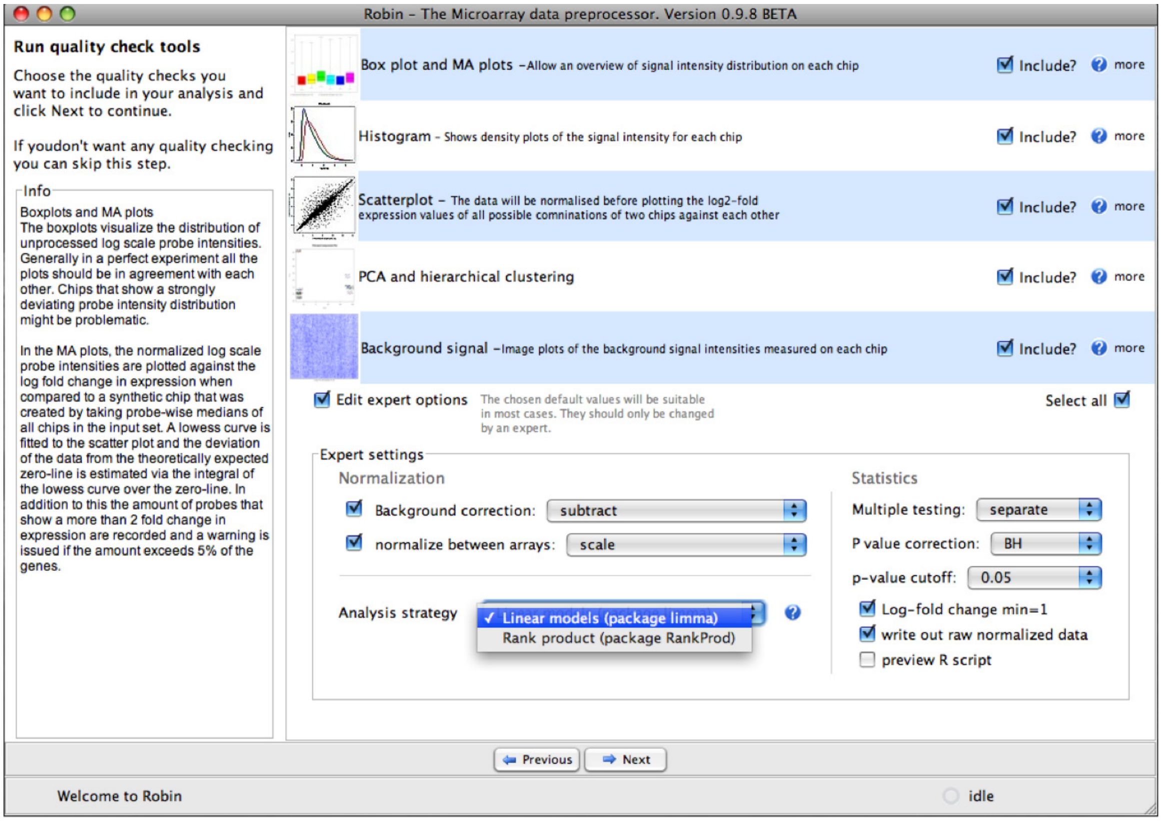
☐ Edit manually

grid rows: 1 block rows: 5

grid columns: 1 block columns: 8

Figure 10: Import dialog for generic single channel microarray data.

Following the review of quality check results as depicted on Figure 4 and described in section 3.1.1, the individual chips have to be organized into groups of biological replicates. Depending on the statistical analysis strategy chosen (rank product- or linear model-based) two or more groups can


Robin - The Microarray data preprocessor. Version 0.9.8 BETA

Run quality check tools

Choose the quality checks you want to include in your analysis and click Next to continue.

If youdon't want any quality checking you can skip this step.

Info

Boxplots and MA plots
The boxplots visualize the distribution of unprocessed log scale probe intensities. Generally in a perfect experiment all the plots should be in agreement with each other. Chips that show a strongly deviating probe intensity distribution might be problematic.

In the MA plots, the normalized log scale probe intensities are plotted against the log fold change in expression when compared to a synthetic chip that was created by taking probe-wise medians of all chips in the input set. A lowess curve is fitted to the scatter plot and the deviation of the data from the theoretically expected zero-line is estimated via the integral of the lowess curve over the zero-line. In addition to this the amount of probes that show a more than 2 fold change in expression are recorded and a warning is issued if the amount exceeds 5% of the genes.

Box plot and MA plots -Allow an overview of signal intensity distribution on each chip ☑ Include? more

Histogram - Shows density plots of the signal intensity for each chip ☑ Include? more

Scatterplot - The data will be normalised before plotting the log2-fold expression values of all possible comnations of two chips against each other ☑ Include? more

PCA and hierarchical clustering ☑ Include? more

Background signal -Image plots of the background signal intensities measured on each chip ☑ Include? more

☑ Edit expert options The chosen default values will be suitable in most cases. They should only be changed by an expert.

Select all ☑

Expert settings

Normalization

☑ Background correction: subtract

☑ normalize between arrays: scale

Analysis strategy ✓ Linear models (package limma) / Rank product (package RankProd)

Statistics

Multiple testing: separate

P value correction: BH

p-value cutoff: 0.05

☑ Log-fold change min=1

☑ write out raw normalized data

□ preview R script

Figure 11: Quality check and expert settings panel for non-Affymetrix single channel microarrays

4 Chip quality assessment

When analysing your own primary microarray data or reanalysing data that is publicly available the first step is quality assessment. Individual chips displaying a very bad quality might strongly impact the final results of your microarray experiment and hence lead to incorrect biological assumptions. Chip quality can be affected on different levels and Robin offers a range of informative plots that cover many different aspects of the chip data quality. In the following section these methods will be described in detail.

4.1 Affymetrix chip quality checks

4.1.1 Analysis of signal intensity distribution

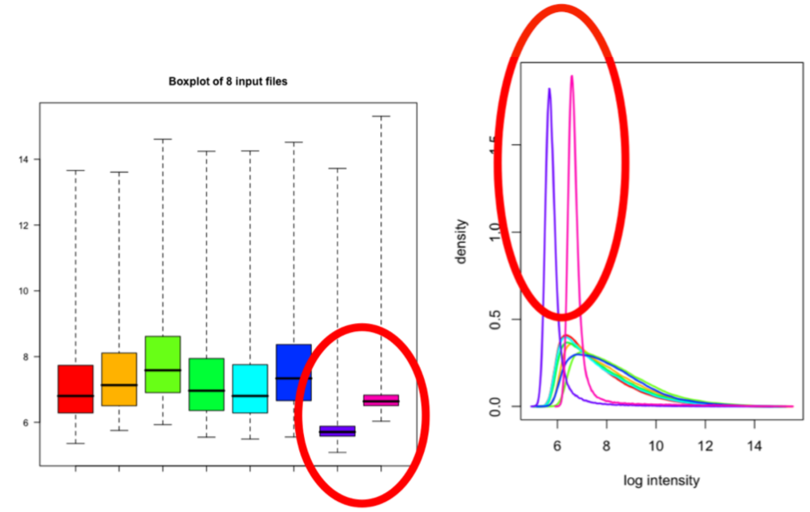


Figure 12: Box plots (left panel) and smoothed signal intensity densities (right panel). The red circles highlight individual chips that show strong outlier behaviour indicating low quality.

Box plots of the unnormalized expression values on each chip give a global overview of the signal intensity distributions. Ideally all chips should have a comparable distribution already before normalization (see Figure 12, left panel). Another way to visualize the distribution of signal intensities is plotting smoothed histograms of the (log2) signal intensity of all perfect match (PM) probes (see Figure 12, right panel). The red circles point out outliers.

4.1.2 MA plots

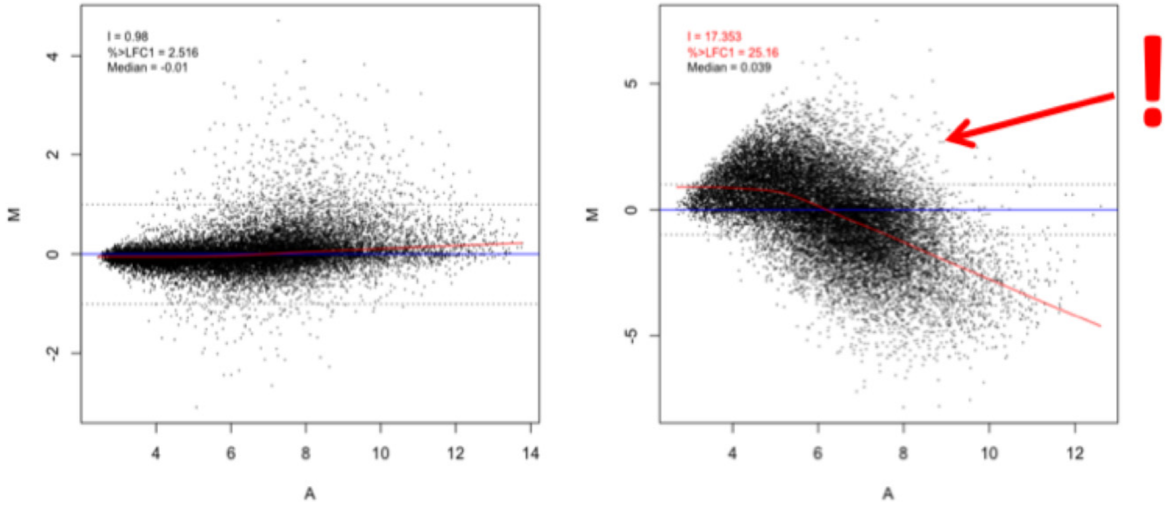


Figure 13: MA plots and box plots. The left panel shows an unobjectionable behaviour while the data displayed on the right panel strongly deviates from normal values. In the box plot (see Figure 12) the two highlighted chips are also clearly showing an outlying intensity distribution.

On the MA plots, the average log2 probe signal intensity **A = ½ * (logR + logG)** is plotted against the log2 fold change in expression **M = logR – logG.** In the case of Affymetrix and other single channel chips G is a synthetic chip created from the median expression values of all chips in the input. For two-color chips the M values are calculated as the log2-fold ratio of the normalized red and green signal intensity. Based on the assumption that most of the genes will not show differential expression. Robin will issue a warning if more than 10% of the genes show a greater than two-fold change (log2 fold change of 1, resp. **M >= 1** or **M <= -1**) in expression. The actual percentage of genes showing a higher than two-fold change in expression is shown on the plot as "%>LFC1". To capture artifacts that are related to the signal intensity **A**, a lowess fit curve over the data points is calculated (see Figure 13). If the integral of the absolute values of the lowess curve over the zero line is greater than 1 another warning is generated indicating that there seems to be a signal intensity-dependent artifactual effect (the integrals value is shown on the plot as "I"). The median **M** value also given on each plot is usually less informative as can be seen on the right panel of Figure 13 where the median shows a normal value while the data quality is severely affected. MA plots are available for all microarray types.

4.1.3 False color images of probe level model weights

A linear model is fitted (using RMA style, more later) to your probeset (i.e. your 11 probes), using the boundary, that the effect of all probes in each probeset is zero.

Weights are attached to the different probes in each probeset, low weights are coloured in green (i.e. they were not important for the model), and high values in white.

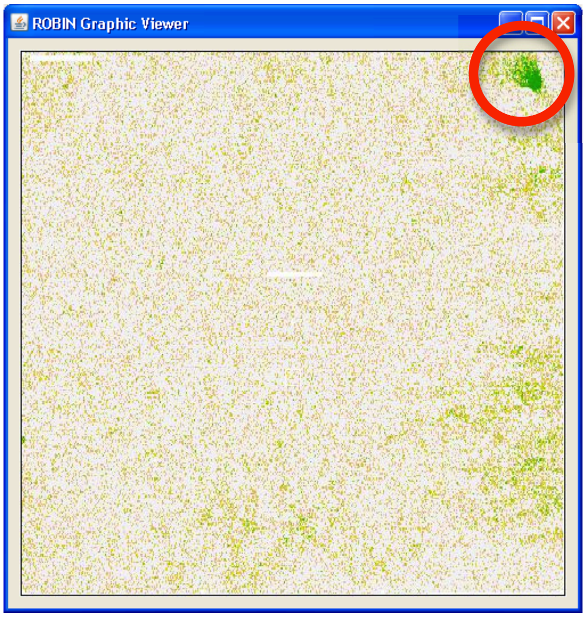


Figure 14: PLM weight image. Here a potential artifact is visible in the upper right corner.

For some examples of probe level model (PLM) image plots showing different artefact have a look at: http://plmimagegallery.bmbolstad.com/. The weights applied to each probe are visualized as pseudo chip images (see Figure 14). Areas on the chip that show consistently low probe weights might indicate technical problems cause e.g. by washing, dust on the chips or scanner malfunction. PLM-based analyses (pseudo images, NUSE and RLE, see next section) are only available for Affymetrix chips.

4.1.4 Normalized unscaled standard error and relative logarithmic expression

The normalized unscaled standard error (NUSE) plots show the standard error estimates of probe level models for each probeset standardized across all chips so that the median standard error for each probeset is 1. The NUSE plot visualizes the distribution of the standard errors for each individual chip. Chips showing a consistently increased standard error are probably of lower quality. The relative logarithmic expression (RLE) is computed by comparing the logarithmic expression of each probeset on each chip to the median expression of this probeset across all chips. According to the assumption that most of the genes are not differentially expressed under a given treatment, the median RLE value should be zero. Individual arrays showing a deviation of the median from the zero line and/or increased spread on a box plot of the RLE values are presumably of low quality.

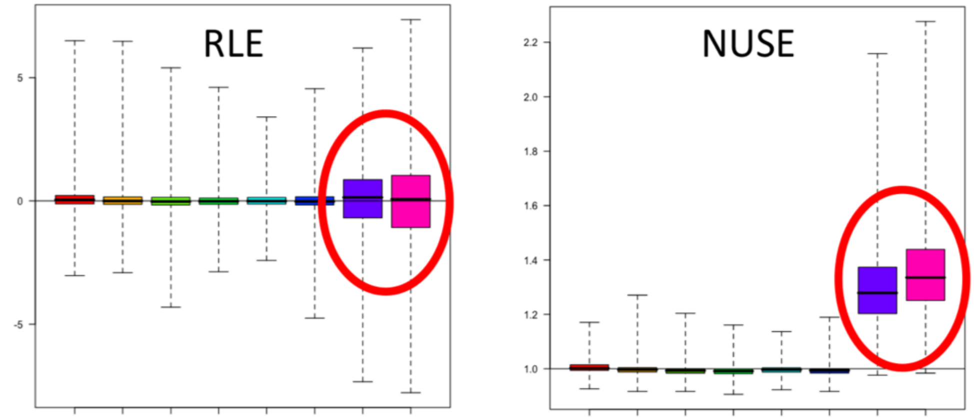


Figure 15: Relative logarithmic expression and normalized unscaled standard error plots. Note the two arrays that are consistently showing low quality behaviour across both plots

4.1.5 RNA degradation

In each probeset the probes are ordered directionally from the 5' to the 3' end. Average probe intensities are plotted by probe number. The resulting plot visualizes the global RNA degradation state of the samples used. Generally, RNA degradation is more active at the 5' terminus - signal intensities of the probes closer to this terminus are accordingly lower. If the slope of the probe intensity curve is exceeding a certain threshold value or the slopes of individual chips are deviating from the median by more than 10% Robin issues a warning (see Figure 16). As this kind of analysis relies on probesets consisting of more than one probe, it is only available for Affymetrix arrays.

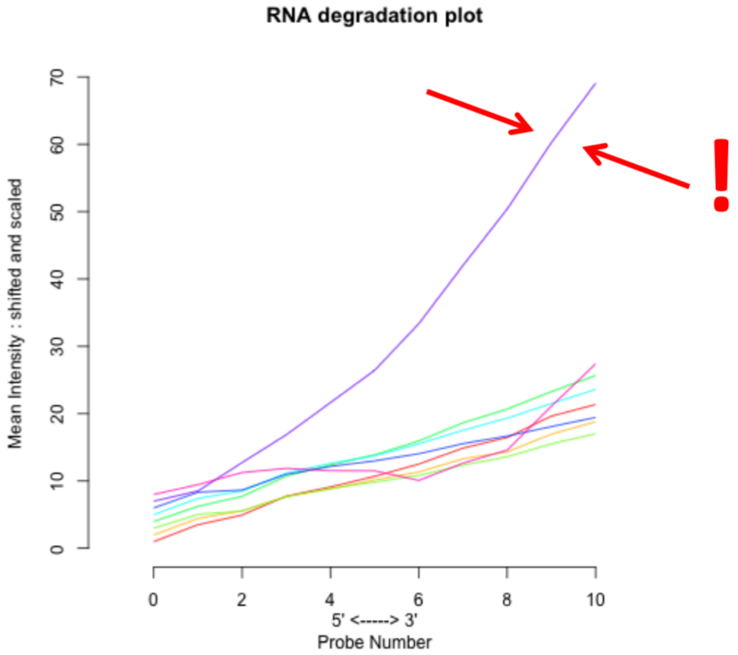


Figure 16: RNA degradation plot.

4.1.6 Scatterplots

If the scatter plot option is chosen, Robin plots pair wise comparisons of the normalized expression values of all possible combinations of two chips. NOTE: Using this feature on a large number of input chips will generate a lot of images and might increase calculation time and memory demand significantly. The scatter plots are useful for assessing whether two replicate chips are showing similar behavior. If they do, the points should lie on a perfect diagonal line. Replicate samples that are not showing this behavior strongly suggest a problem (e.g. accidentally swapped or mislabeled samples, technical problems on one individual chip, strong RNA degradation effects etc.)

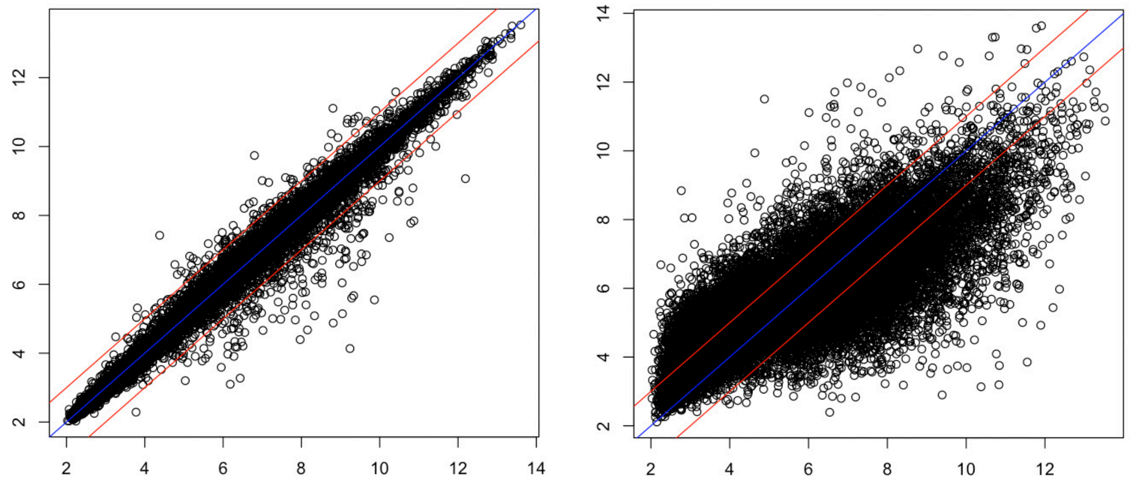


Figure 17: Scatter plots of normalized expression values. The left panel shows two biological replicates of acceptable reproducibility plotted against each other while the right panel shows two chips with very different expression profiles. Identical values are plotted on the blue (0) line; The red lines indicate a log2-fold difference of 1.

4.1.7 Principal component analysis and hierarchical clustering

The data generated in a microarray experiment can be understood as a matrix of ***p*** columns, where ***p*** is the number of chips used, and ***n*** rows, where ***n*** is the number of genes (probesets, probes) measured. Such a dataset could be visualized as a set of ***n*** points in a ***p-dimensional*** space. The principal component analysis reduces the dimensionality of the dataset by finding a small number of linear combinations of the data that explain most of the variance in the dataset. These are the principal components (PCs). The principal components are ordered by the amount of variance explained and subsequently the first two PCs are plotted against each other. The example on the right panel of Figure 18 shows a PCA on eight samples, six of which are grouping closely together on two groups of three replicates while the last two are completely unrelated.

The hierarchical clustering method performs a clustering of the Pearson correlation of raw normalized expression estimates for each chip a the distance measure for the clustering. Chips that show similar expression profiles should cluster together when using this approach. The results are shown as a dendrogram where the branch length depicts 1-correlation score. The hierarchical clustering gives an overview of the internal structure of the data and identifies experimental conditions that generate similar global responses in gene expression. Replicate chips should always cluster closely. Accordingly, the samples six samples belonging to two groups of three replicates form distinct clusters while the last two are very distant from them and each other. The PCA and hierarchical clustering analyses are only available for Affymetrix and generic single channel experiments.

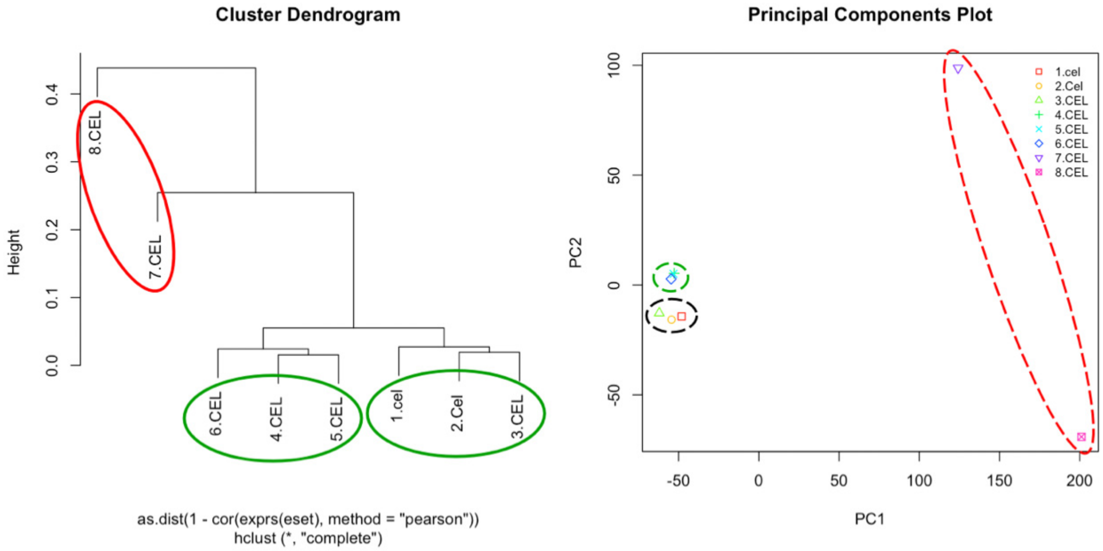


Figure 18: Principal component analysis and hierarchical clustering of normalized expression values. The red circles highlight chips with strongly deviating behavior.

4.2 *Two color microarray quality checks*

Quality check methods that are specific to two color arrays are described in the following section. Some quality checks that can be run for all chip types – these will not be described again below (e.g. MA plots).

4.2.1 Image plots of two-color background intensities and unnormalized M values

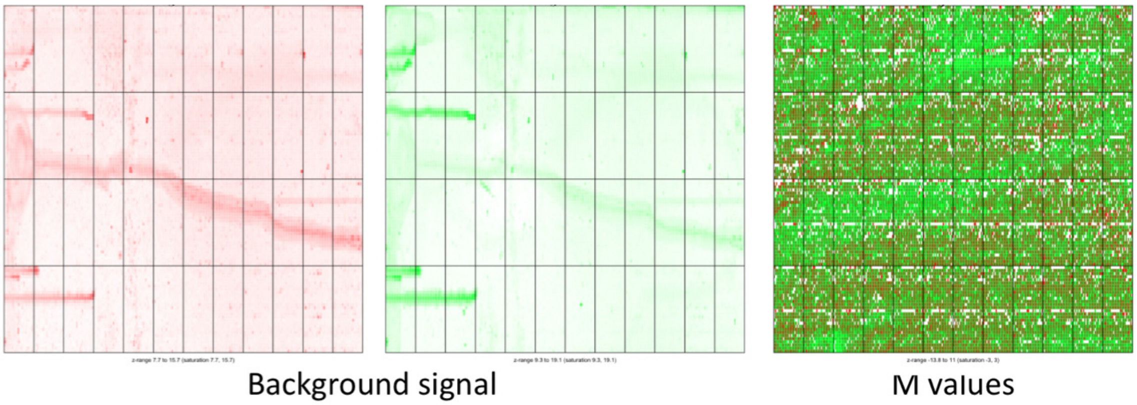


Figure 19: Two-color microarray background signal intensities and unnormalized M value plots.

The background signal intensities measured on two-color and generic single channel chips (not shown here) can be visualized as false-color images. This is very useful for the identification of washing artifacts like those visible on the two left plots in Figure 19. Both color channels display obvious traces of droplets, so called washing artifacts. In the worst case these artifacts carry over to the foreground signal and cannot be eliminated by background subtraction. If this happens they would also be visible on the M value plot shown on the right side of Figure 19 (in the example given, however, this is not the case). The M value plots is simply a false-color image of the merged red and green foreground signal intensities measured on the chip prior to normalization.

4.2.2 Overview of two color signal intensity distribution

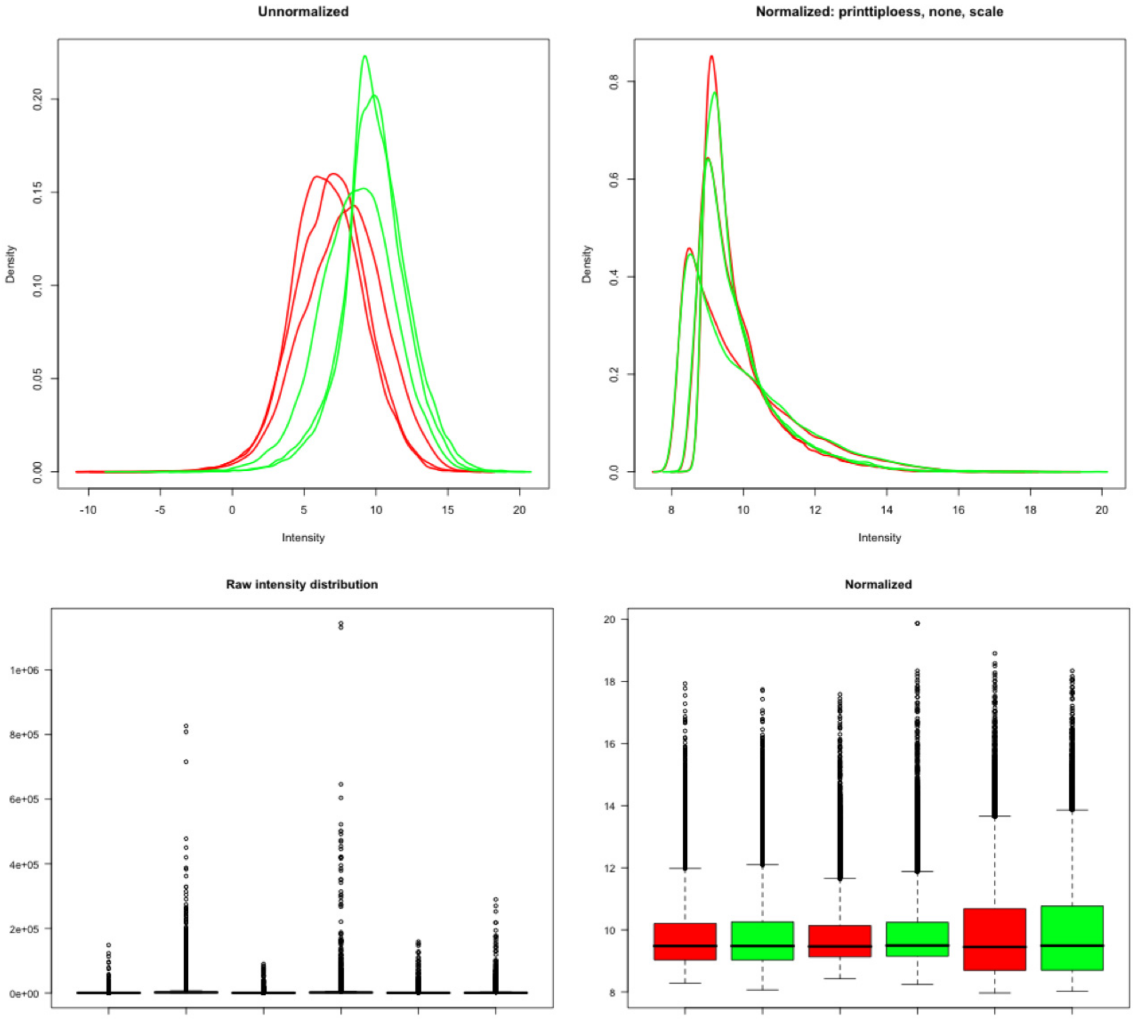


Figure 20: Two color microarray signal intensity distribution assessment. Upper left: Smoothed signal intensity distributions are shown for the red and green channel separately for each chip. Lower left: Box plots of the raw foreground signal intensities for each chip and color channel. The left hand plots show data prior to normalization while the plots on the right half show normalized data. The title of the right hand intensity distribution plot reflects the chosen normalization settings. For the shown example within-array printtiploess normalization without background correction and between-array scaling were performed.

Analogous to the box plots and smoothed histograms that are generated for Affymetrix arrays (see section "Analysis of signal intensity distribution" and Figure 12).

5 Data normalization

When analyzing microarray experiments, the raw data obtained by scanning probe intensities on the chips can be strongly influenced by different technical effects. These can be different levels of background signal due to inhomogeneous washing, systematically deviating probe signal intensities due to different scanner settings (or even same settings on different devices), probe-specific hybridization affinity effects etc.
To make sure that the microarrays you are going to analyze in a differential expression experiment can actually be compared it is very important to eliminate these effects. This process is called normalization. Since the first application of microarray technology many different normalization techniques have been proposed - the most widely used ones are available in Robin. If your favorite method is not among them feel free to contact us.

Generally, all normalization methods consist of two (three in the case of Affymetrix GeneChip microarrays) major steps: (I) background correction, (II) normalization of background-corrected probe level data and (III) summarization of probe-level data to yield one expression measure per probeset.

5.1 Single channel microarray normalization

5.1.1 Normalization methods for Affymetrix arrays

5.1.1.1 RMA (Irizarry et al., 2003)

The robust multi-array average (RMA) normalization method proposed by (Irizarry et al., 2003) has been widely used and accepted as a well-performing approach for inference of differential gene expression from Affymetrix GeneChip(R)-based experiments. The RMA procedure first does background correction based on the assumption that the background signal is normally distributed while the real probe signal is exponentially distributed (convolution model). The background-corrected data is then quantile normalized. Quantile normalization assumes that the distribution of gene abundances is nearly the same across all chips. A reference distribution is created using the pooled intensity probe distribution on all chips. To normalize each chip, the quantile of each intensity value is computed and then the original value is transformed to the corresponding quantile's value on the pooled reference chip (that is created by averaging the values of each probe across all chips in the experiment). In the last step, a linear model is fitted to the corrected, normalized and log2-transformed probe intensities: $Y_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn}, i = 1,...,I, n = 1,...,n$ with α_j being a probe affinity effect, μ_i representing the log2 expression level on array I and ε_{ij} representing a noise error with mean = 0. The model parameters are estimated using the median polish procedure that is robust against outliers.

5.1.1.2 GCRMA (Wu et al., 2004)

The GCRMA method adds a more refined background adjustment to the standard RMA normalization. This background adjustment method models the different hybridization affinities for each PM-MM probe pair based on its nucleotide sequence which results in a more precise estimate of the background. While the standard RMA approach ignores the MM probe-derived signal, GCRMA subtracts a shrunken MM value that was corrected for its binding affinity from the PM signal. More specifically, the model assumes: $PM = O_{PM} + N_{PM} + S$ and $MM = O_{MM} + N_{MM} + \phi S$ with O being the optical noise, N being the non-specific binding effect and S being proportional to the real concentration of the target transcript. Hence, the model takes into account the observation, that the MM signal may contain real transcript signal.

5.1.1.3 MAS 5.0 (Affymetrix Microarray Analysis Suite 5.0)

In contrast to the other normalization methods described here, MAS 5.0 works on a single chip basis. Briefly, each chip is divided into 16 (4x4) equally sized grid regions and a background and noise signal value is calculated based on the lowest 2% of measured probe intensities for each grid region. The probe intensities in each grid block are adjusted to the weighted average of the background signal where the weight is dependent on the (euclidean) distance of the probe to the centroid of the grid block. In the next step the perfect match (PM) and mismatch (MM) probe pairs are considered. The original purpose of the PM/MM probe pair design was to use the MM probe signal intensity as unspecific signal intensity and subtract it from the PM probe to generate a reliable probe signal. However it turned out that up to 30% of the MM probes display a signal intensity that is higher than the corresponding PM probe so that a simple subtraction would yield negative values. To work around this problem, the so called ideal mismatch (IM) was introduced. If the PM intensity is larger than MM, IM equals the MM value. In cases where PM=MM or PM<MM, IM is calculated using the PM value and a specific background (SB) value that is computed by taking a robust average of the log ratios of PM and MM. The summarized expression measure is computed using a Tukey biweight of PM and IM values in each probe set on the $\log_2$ scale. In MAS 5.0, the normalization is performed after summarization. A scaling normalization is used to adjust intensity values on each array. MAS 5.0 provides final expression values on the original scale. The Robin analysis workflow takes this into account and logarithmizes the values prior to statistical analysis to provide uniform output independent of the normalization method chosen. For a more detailed description of the method, please see the Affymetrix technical documentation (Affymetrix GeneChip® Expression Analysis, 2004).

5.1.1.4 PLIER (Affymetrix, Probe Logarithmic Intensity Error Estimation, 2005)

The PLIER method was developed by Affymetrix as an improved estimator of signal intensity. It is, unlike MAS 5.0, a multi-array method but includes the summarization algorithm that is also used in the MAS 5.0 method. Like RMA it uses a global model but bases this on a different set of assumptions. Unlike RMA it takes the MM probe signal into account when computing expression values. The observed PM and MM probe signal intensities for the *i*th probe on the *j*th array are assumed to be $E(PM_{ij}) = \mu_{ij} = a_i c_j + B_{ij}$ and $E(MM_{ij}) = B_{ij}$ with μ_{ij} being the binding level of probe i and array j, a_i being the probe specific binding affinity, c_j the RNA concentration in the sample hybridized to array j and B_{ij} the background binding intensity of probe I on array j.

PLIER also assumes that the error of the PM and MM probe signals are reciprocal (while MAS 5.0 assumes them to be equal): $\varepsilon_{ij}^{PM} = \frac{1}{\varepsilon_{ij}^{MM}}$ with ε_{ij}^{PM} being the error of the *i*th perfect match probe in the *j*th array and ε_{ij}^{MM} the error of the corresponding mismatch probe. This results in the following equation: $\varepsilon_{ij} = \frac{a_i c_j + \sqrt{(a_i c_j) + 4 pm_{ij} mm_{ij}}}{2 pm_{ij}}$.

Based on the above assumptions, the PLIER algorithm computes the values of a and c by setting the residual $r = \log(\varepsilon)$ to zero using a minimization of a robust average of the r^2 values. PLIER performs slightly better than MAS 5.0 when comparing the analysis of spike-in experiments where RNA of know concentration was added to the sample, possibly due to a better error estimation procedure. For further details and an in-depth discussion of the PLIER algorithms and performance, please see Affymetrix's Guide to PLIER esitimation, 2005 and Therneau and Ballmann, 2008.

5.1.2 Normalization of generic single channel and two color arrays

Since most of the non-Affymetrix microarrays do not adopt a probeset design where multiple probes are matching one target transcript, the summarization step necessary for Affymetrix raw data is omitted. The two remaining steps, background correction and normalization, can be flexibly configured according to the experiments' requirements and users' preferences.

5.1.2.1 Background correction

Several methods to correct the measured probe intensity for background signal intensity are available. The background signal intensity values themselves have to be provided in a separate column in the raw data file and have to be specified upon import of the data (please see 3.2 and 3.3). Aside from "subtract" all background correction procedures are designed to produce positive corrected signal intensities.

1) "subtract"– Simply subtracts the background intensities from the foreground intensity values.

2) "half" – All foreground signal intensities that are less than 0.5 of the original intensity after background subtraction will be set to 0.5 of the uncorrected value.

3) "minimum" – Values that are zero or negative after simple background subtraction are set to 0.5 times the smallest positive corrected value.

4) "edwards" – Uses a log-linear interpolation to adjust low intensity values (see Edwards, 2003)

5) "normexp" – Uses the same convolution model that is applied in the RMA method to model the background intensity with two modifications to make it better applicable for two color arrays. First, the model is fitted to the background subtracted foreground values of each color channel separately and second, instead

of using a kernel density parameter estimator for the model parameters, a maximum-likelihood estimator is used See (Ritchie et al., 2007) for details.

6) "rma" – Employs the background correction step of the RMA method for Affymetrix arrays.

5.1.2.2 Within-array normalization

This option is only available for two color microarrays and normalizes the log2 ratios of expression of the red and green channel signals so that the average log2-ratio is zero. This is again based on the assumption that most of the genes do not show differential expression in a given experiment. The options available will be described in the following:

1) "median" – Simply subtracts the median from the calculated M values.

2) "loess" – Uses global loess regression (a robust smoothing algorithm based on local polynomial regression) to compute a trend in the data. Each M value is normalized by subtracting the corresponding the corresponding value of the loess curve from it according to $N = M - loess\ (A)$, where N is the normalized value, M the raw value and *loess(A)* the loess curve as a function of the average signal intensity A.

3) "printtiploess" – Performs the loess normalization separately for each print tip group. This approach accounts better for local spatial variation in background signal intensity and it therefore used as the default method for within-array normalization in Robin.

4) "robustspline" – This method does also normalize print tip group-wise but uses regression splines and empirical Bayes-based shrinkage instead of loess curves for normalization.

5.1.2.3 Between-array normalization

In addition to normalizing within each two color array, the user can choose to also perform between array normalization. When analyzing single channel arrays, this is the only normalization approach available. Applying between array normalization makes sure that the expression intensities (resp. log2 ratios on two color arrays) have equal distributions across a series of chips. Several options are available for two color arrays while the list is limited to scale and quantile normalization for single channel arrays.

1) "scale" – Log2 ratios of expression are scaled to have the same median absolute deviation (MAD) across arrays.

2) “quantile” – Adjusts to intensities to have the same empirical distribution across chips. This is the normalization method that is also used by the RMA procedure for Affymetrix chips.

3) “Aquantile” – Is a variation of the quantile method that only adjusts the A values to display the same distribution.

4) “Tquantile” – Does a quantile normalization separately for each of the target groups defined on the targets designer panel in the two color chip analysis workflow (see 3.2).

5) “Gquantile” and “Rquantile” – Quantile normalization is performed for the green (“G”) or red (“R”) color channel only. This approach makes sense if a common reference design has been employed in the experiment that is being analyzed and the reference sample was always hybridized in the same color channel.

Both normalization approaches can be combined when working with two color channels. In this case, within array normalization and background correction are performed prior to between array normalization steps. The preset default settings should give robust expression estimates in most cases. However, given the heterogeneity of two color and single color technical chip platforms, different settings may perform better for individual chip types. When trying to assess whether the chosen settings give decent results in a given experiment, it helps to inspect the shape of the MA plots after normalization. If the distribution of values displays the expected (often “trumpet”-like) shape and the plot is centered on the M = 0 line (see Figure 13), the settings seem to be sound. If in doubt, please seek advice from an experienced statistician.

6 Analysis of differential gene expression

The statistical methods Robin employs to identify differentially expressed genes are based on two different approaches: Linear modeling (limma, (Smyth, 2004)) and rank product-based analysis (RankProd, (Breitling et al., 2004; Hong et al., 2006)). When analyzing Affymetrix data, the user can choose between these two options with the restriction that rank product-based inference of differential expression is only available when two groups are to be compared. When working with two color microarrays, rank product-based analysis is not available yet. The two methods differ in that they take two completely different approaches to the detection of differentially expressed genes. While the linear model-based method relies on advanced statistical modelling and Bayesian inference, the rank product approach more resembles biological reasoning on the data. More specifically, limma assumes a linear model $E\left[y_j\right] = X\alpha_j$ where y_j contains the expression data for each gene, X is the design matrix describing the systematic part of the data and α_j is a vector of coefficients (representing the response level for gene j on chip g). The biologically interesting contrasts of the coefficients are defined by $\beta_j = C^T\alpha_j$, where C is the contrast matrix (for a more detailed in-depth discussion please refer to (Smyth, 2004)).

The rank product approach, on the other hand, assumes for an experiment in which n genes are investigated in k replicates, that the probability to find a gene at the top position of a ranked list of up- or down regulated genes is exactly $1/n^k$. The combined probability of finding a gene at a certain position in the ranked list, when k replicates i and n_i genes are measured can be expressed as the rank product $RP_g^{up/down} = \Pi_{i=1}^{k}(r_{i,g}^{up/down}/n_i)$, where $r_{i,g}^{up/down}$ is the position of gene g in the ranked list of decreasing (*up*) or increasing (*down*) fold changes in the *i*th replicate (see (Breitling et al., 2004) for further details not to be reproduced here).
Since rank product-based analysis is limited to comparing two experimental conditions, the linear model-based analysis offers far more options and flexibility with respect to the available settings and design of the experiment (e.g. if two factors, like genotype and treatment, are being varied in an experiment and the user is interested in the interaction effect).

7 Output

At the end of each analysis run, Robin asks for a directory to save all files that are relevant to the experiment. These include processed raw input data files (only in the case of two color and generic single channel analysis), R source code for quality assessment and main analysis, various informative plots illustrating the quality check results and main analysis results and tabular text data files containing the full results in all detail. The following table lists all files that are generated. The "Type" column refers to the microarray type for which this kind of output file can be generated (G = all platforms, A=Affymetrix, T=two color microarrays, S=generic single channel arrays). Parts of the file names written in italics refer to variable text: *EXP_NAME*: The name of the experiment as entered by the user when choosing the name of the output folder. *TMP*: An automatically generated unique identifier used for temporary files (the quality check output files are first stored in the system's temporary folder and are later copied to the quality checks folder of the output directory). *GRP*: Reference to the group names as assigned by the user when sorting individual raw files (e.g. .cel files) into groups of biological replicates.

Filename	**Folder**	**Description**	**Type**
*EXP_NAME*_results.txt	.	This file contains the normalized log2 fold change in expression values for all comparisons defined in the design step of the experiment. In addition, a second column containing a flag value denoting the statistical significance of each log fold change is generated for each gene. A value of 0 means not significant, while -1 and 1 mean significantly up- or down-regulated.	G

*EXP_NAME*_summary.txt	.	A text file summarizing and documenting the analysis inputs, program settings and warnings generated during the workflow.	G
*EXP_NAME*_design.png	.	PNG representation of the experiment design as configured on the graphical designer panel in the last step of the analysis workflow.	G
redundant.probes.info.txt	detailed_results	If redundant probe names are found in the input data of the generic single channel rank product analysis, this file is generated. It contains the redundant identifiers, number of spots found and the median values for each of the identifiers on each chip.	S
full_table_*GRPa-GRPb*.txt	detailed_results	Tables giving the complete statistical results for each of the comparisons made. The columns contain from left to right: (Feature.ID) A unique identifier for the oligonucleotide probes or probe sets on the chips; (logFC) the log2-fold change in expression; (AveExpr) average normalized expression value; (t) t-statistic; (P.Value, adj.P.Val) raw and Benjamini-Hochberg-corrected p-values for differential expression; (B) the log-odds for differential expression.	G
top100table_*GRPa-GRPb*.txt	detailed_results	Contains the same data columns as the full tables but excludes probes / probesets that do not fulfill the chosen p-value and or minimal log2-fold change cut offs.	
EXP_NAME.PAcalls.table .txt	detailed_results	Only generated when analyzing Affymetrix chips. Table containing the present / absent calls for each probeset on each chip in the experiment plus the attached p-values that are calculated using the MAS5calls function.	A
raw_*METHOD*_normalized_expression_values.txt	.	Expression estimates for each probe/probeset on each chip after normalization.	A
*TMP*_hclust.png	qualitychecks	Hierarchical clustering of the normalized expression values. The clustering is based on 1-pearson correlation of expression as the distance measure. Full linkage hierarchical clustering is performed.	A, S
*TMP*_pcaplot.png	qualitychecks	Scatter plot of the first two components obtained in a principal component analysis of the normalized expression values.	A, S

*TMP*_boxplot.png	qualitychecks	Boxplots of the unnormalized signal intensities on each chip	G
*TMP*_hist.png	qualitychecks	Smoothed density plots showing the signal intensity distribution on each chip prior to normalization.	A, S
*TMP*_density.png	qualitychecks	These plots display the signal intensity distribution for two color arrays analogous to the ”hist” plots for Affymetrix and other single channel arrays. Smoothed distributions are plotted separately for both color channels	T
*TMP*_maplot*1..n*.png	qualitychecks	MA plots of chip 1 to n. When analyzing single channel chips, these plots show the log2-fold change in expression of each individual chip when compared to a synthetic chip constructed from the median expression values of all chips in the experiment. In the case of two color arrays the M values correspond to log log2 ratio between the green and red channel signal intensities prior to and after normalization. Each plot also shows the following quality-associated parameters: “I” – Absolute value of the numerical integral of the lowess fit curve over the M=0 line. Values greater than 1 are considered to indicate lower quality. “%>LFC1” - Percentage of probes/probesets displaying a log2-fold change greater than 1. Based on the assumption that most of the genes will not show differential expression, a warning will be issued of more than 5% of the probes show an absolute log2 fold change higher than 1 (meaning 2-fold up- or downregulation). “median” – Gives the median value of M. In an ideal experiment this should be zero.	G
*TMP*_plm*1..n*.png	qualitychecks	Shows pseudo images of the model weights for each probe after fitting linear probe level models. Low weights are indicated by stronger red or green color	A
*TMP*_rle.png	qualitychecks	Boxplots of the relative logarithmic expression (RLE) values on each chip. The boxes should be centered around zero.	A

*TMP*_nuse.png	qualitychecks	Boxplots of the normalized unscaled standard errors (NUSE) of the probe level models on each chip. The plots should be centered around zero and display comparable spread.	A
*TMP*_scat*1..n1..m*.png	qualitychecks	Scatter plots of all possible combinations of two chips. The normalized expression values are plotted against each other.	A, S
*TMP*_rna.png	qualitychecks	RNA degradation plot (only available for Affymetrix arrays). Shows mean intensities of probes in all probesets ordered from the 5' to the 3' end of the target sequence. This plot allows a good overview of the global RNA quality on the chips.	A
*TMP*_bground.png	qualitychecks	Pseudo images of the background signal intensities measured on two color or non-Affymetrix single channel arrays.	T, S
*TMP*_mvalues.png	qualitychecks	Pseudo image plots of the unnormalized M (= log2 ratios of green and red signal) values of two color chips.	T
*XYZ*_robin	input	Cleaned-up copies of the input raw data files	T, S
EXP_NAME.main.analysis .R	source	The R script file containing code for the main analysis. The file can be used as a starting point for customizations of the analysis. Please note that the file contains some hard coded paths.	G
qualityChecks.R	source	Quality checks R source code file.	G
MAplot_*GRPa-GRPb*.png	plots	The plots folder contains some informative plots on the results of the main analysis. MA plots are generated for each contrast that was defined on the experiment designer panel. Genes that are significantly differentially expressed according to the statistical analysis are highlighted by red circles.	G
vennDiagram_down/total/ up.png	plots	Venn diagrams showing the number of significantly up- down- and total regulated genes for up to four contrasts.	G
PCAplot.png	plots	Principal component analysis plot analogous to the plots generated in the quality checks section. This plot does in addition highlight the groups of replicate experiments as defined on the groups panel.	A, S

REFERENCES

Breitling R, Armengaud P, Amtmann A, Herzyk P (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett **573:** 83-92

Holm S (1979) A stagewise rejective multiple test procedure based on a modified Bonferroni test. Scandinavian Journal of Statistics**:** 65-70

Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics **22:** 2825-2827

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England) **4:** 249-264

Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK (2007) A comparison of background correction methods for two-colour microarrays. Bioinformatics **23:** 2700-2707

Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology **3:** Article3

Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. Journal of the American Statistical Association **99:** 909-917